

WAVELET PENALIZED LIKELIHOOD ESTIMATION IN GENERALIZED PARTIALLY LINEAR MODELS

Irène Gannaz

Université de Lyon

CNRS UMR 5208

INSA de Lyon

Institut Camille Jordan

20, avenue Albert Einstein

69621 Villeurbanne Cedex

France.

E-mail: irene.gannaz@insa-lyon.fr

Abstract

The paper deals with a semiparametric generalized partially linear regression model with unknown regression coefficients and an unknown nonparametric function. We present a maximum penalized likelihood procedure to estimate the components of the partial linear model introducing penalty based wavelet estimators. Asymptotic rates of the estimates of both the parametric and the nonparametric part of the model are given and quasi-minimax optimality is obtained under usual conditions in literature. We establish in particular that the ℓ^1 -penalty leads to an adaptive estimation. An algorithm is also proposed for implementation and simulations are used to illustrate the results.

Keywords: Semiparametric models, generalized partially linear models, M-estimation, penalized loglikelihood estimation, wavelet thresholding, backfitting.

1 Introduction

The aim of this paper is to consider a regression model, where the response Y is to be predicted by covariates \mathbf{z} , with Y real-valued and with \mathbf{z} a real explanatory vector. Relaxing the usual assumption of normality, we consider a generalized framework. The response value y is drawn from a one-parameter exponential family of distributions, with a probabilistic density of the form:

$$\exp \left(\frac{y\eta(\mathbf{z}) - b(\eta(\mathbf{z}))}{\phi} + c(y, \phi) \right). \quad (1)$$

In this expression, $b(\cdot)$ and $c(\cdot)$ are known functions, which determine the specific form of the distribution. The parameter ϕ is a dispersion parameter and is also supposed to be known in what follows. The unknown function $\eta(\cdot)$ is the natural parameter of the exponential family, which carries information from the explanatory variables. Given a random sample of size n drawn independently from a generalized regression model, the aim is to predict the function $\eta(\cdot)$. Such a model gives a large scope of applications because observations can result from many distribution families such as Gaussian, Poisson, Binomial, Gamma, *etc.* For a more thorough description of generalized regression modelling, we refer to McCullagh & Nelder (1989) or Fahrmeir & Tutz (1994).

Let $(Y_i, \mathbf{z}_i)_{i=1, \dots, n}$ be an independent random sample drawn from a generalized regression model. The conditional mean and variance of the i^{th} response Y_i are given by:

$$\mathbb{E}[Y_i | \mathbf{z}_i] = \dot{b}(\eta(\mathbf{z}_i)) = \mu(\mathbf{z}_i), \quad (2)$$

$$\text{Var}[Y_i | \mathbf{z}_i] = \phi \ddot{b}(\eta(\mathbf{z}_i)), \quad (3)$$

where $\dot{b}(\cdot)$ and $\ddot{b}(\cdot)$ denote respectively the first and second derivatives of $b(\cdot)$. The function $G = \dot{b}^{-1}$ is called link function and one have $G(\mathbb{E}(Y_i | \mathbf{z}_i)) = \eta(\mathbf{z}_i)$.

A linear model consists in assuming that the dependence from the covariate \mathbf{z} is linear, meaning $\eta(\mathbf{z})$ can be written on the form $\eta(\mathbf{z}) = \mathbf{z}^T \boldsymbol{\beta}$; the superscript T denotes the transpose of a vector or matrix. Yet, in some applications, the linear model is insufficient to explain the relationship between the response variable and its associate predictors. The generalized functional model relax this linearity, considering a nonparametric form for the canonical parameter, say $\eta(\mathbf{z}) = f(\mathbf{z})$.

Nevertheless in such a model appears the well-known *curse of dimensionality*. To avoid this drawback, we allow most predictors to be modelled linearly, while a small number is modelled nonparametrically. To this aim we decompose the covariate \mathbf{z} in two components: in the following $\mathbf{z} = (\mathbf{X}, t)$, with \mathbf{X} a p -dimensional vector and t a real-valued covariate. The function $\eta(\cdot)$ is given by:

$$G(\mu(\mathbf{X}, t)) = \eta(\mathbf{X}, t) = \mathbf{X}^T \boldsymbol{\beta} + f(t), \quad (4)$$

where $\boldsymbol{\beta}$ is an unknown p -dimensional real parameter vector and $f(\cdot)$ is an unknown real-valued function; such a model is called a generalized partially linear model (GPLM). Given the observed data $(Y_i, \mathbf{X}_i, t_i)_{i=1, \dots, n}$, the aim is then to estimate from the data the vector $\boldsymbol{\beta}$ and the function $f(\cdot)$.

In a Gaussian modelisation, Rice (1986) and Speckman (1988) put in evidence that the rates of the estimates for linear and nonlinear parts could not be both optimized without a control of the correlation between the explanatory variable of the linear part and the functional part of the model. With such a control, Speckman (1988) proves that it is possible to obtain both optimal linear rate and nonparametric rate for the estimates. To my knowledge, the only paper establishing such a result in GPLM is Mammen & Van der Geer (1997).

Many papers focus on the asymptotic behaviour of the estimator of the parametric part $\boldsymbol{\beta}$ in generalized partially linear models (see *e.g.* Chen (1987) by a penalized *quasi-least squares* or Severini & Staniswalis (1994) by *profile likelihood* methods). A recent article of Boente et al. (2006) establishes a uniformly convergent estimation for f using a robust *profile likelihood* similar to Severini & Staniswalis (1994). But few works consider simultaneously the parametric and the nonparametric part of the model. The paper of Mammen & Van der Geer (1997) shows minimax optimality for the estimations of both f and $\boldsymbol{\beta}$ with a penalized *quasi-least squares* procedure. Note that the authors use a Sobolev type penalty. The conditions given there for attaining optimality for both parametric and nonparametric estimators appear to be more restrictive than those given in Gaussian partially linear models in the literature.

This paper proposes a new estimation procedure based on wavelet decompositions. Wavelet based estimators for the nonparametric component of a Gaussian partially linear model have been investigated by F. Meyer (2003), Chang & Qu (2004), Fadili & Bullmore (2005) or Gannaz (2007b)

and Gannaz (2007a) more recently. But it has not been studied in the context of GPLM. Estimation methods encountered in literature need the choice of a smoothing parameter, which optimal value depends on the regularity of the functional part f . A cross-validation procedure is then necessary to evaluate this parameter. As noted by Rice (1986) or Speckman (1988), cross-validation can present much instability in partially linear models. The use of wavelets here leads to a procedure where no cross-validation is needed. This adaptivity is the main novelty of our estimation scheme in such models. We moreover establish the near-minimax optimality of the estimation for both the linear predictor β and the nonparametric part f , under usual assumptions of correlation between the two parts. The correlation condition appears to be similar to what is classically for the Gaussian case and is weaker than in Mammen & Van der Geer (1997). Finally, we present an algorithm for computing the estimates.

The paper is organized as follows: Section 2 presents the assumptions and the estimation procedure. It also gives the main properties of our estimators. We distinguish two cases: non adaptive penalties and a ℓ^1 type penalty, which leads to adaptivity. In Section 3, we propose a computational algorithm of the adaptive estimation procedure and we present a small simulation study for the numerical implementation. Proofs of our results are given in the Appendix.

2 Assumptions and estimation scheme

We consider a generalized regression model, where the response Y depends on covariates (\mathbf{X}, t) , where Y is real-valued, \mathbf{X} is a p -dimensional vector and t is a real-valued covariate. The value y is drawn from an exponential family of distributions, with a probabilistic density of the form:

$$\exp \left(\frac{y\eta(\mathbf{X}, t) - b(\eta(\mathbf{X}, t))}{\phi} + c(y, \phi) \right),$$

where functions $b(\cdot)$ and $c(\cdot)$ as well as the dispersion parameter ϕ are supposed to be known. The aim is to evaluate the unknown generating function $\eta(\cdot)$. As noted above, we are interested in this paper by a partially linear modelisation of the function $\eta(\cdot)$ and hence we suppose that it has the semiparametric expression:

$$\eta(\mathbf{X}, t) = \mathbf{X}^T \beta + f(t).$$

The vector β and the function f are respectively the parametric and nonparametric components of the generalized partially linear model (GPLM). In the following, $(Y_i, \mathbf{X}_i, t_i)_{i=1, \dots, n}$ will denote an independent random sample drawn from the GPLM described here.

2.1 Penalized maximum likelihood

The aim of the paper is to estimate simultaneously the parameter β and the function f , given the observed data $(Y_i, \mathbf{X}_i, t_i)_{i=1, \dots, n}$. We propose a penalized maximum loglikelihood estimation. Let ℓ denotes the loglikelihood function: $\ell(y, \eta) = y\eta - b(\eta)$. We consider throughout the paper that the estimators \hat{f}_n and $\hat{\beta}_n$ are solutions of :

$$(\hat{f}_n, \hat{\beta}_n) = \underset{\{f, \|f\|_\infty \leq C_\infty\}, \beta \in \mathbb{R}^p}{\operatorname{argmax}} K_n(f, \beta) \quad \text{with} \quad K_n(f, \beta) = \sum_{i=1}^n \ell(y_i, \mathbf{X}_i^T \beta + f(t_i)) - \operatorname{Pen}(f). \quad (5)$$

We refer to Antoniadis et al. (2009) for the conditions of existence of such a maximization problem. In what follows, we will assume the penalty is convex and the likelihood function is bounded in order to ensure the existence of maxima (unicity is not acquired but there is no local maxima). We did not succeed in getting rid of the constraint $\|f\|_\infty \leq C_\infty$ in the proofs but such a condition does not seem too restrictive in practice.

The computation of the maximization problem is done in two step. Some studies, such as Speckman (1988), incite to estimate first the functional part. Thus, we will proceed as follows:

1. $\tilde{f}_n, \beta = \underset{\{f, \|f\|_\infty \leq C_\infty\}}{\operatorname{argmax}} K_n(f, \beta)$.
2. $\hat{\beta}_n = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}} K_n(\tilde{f}_n, \beta)$. Actually, a classical procedure is here to maximise a modified criterion called *profile likelihood* (see among others Severini & Wong (1992) or Boente et al. (2006)). The criterion maximized is then $\sum_{i=1}^n \ell(y_i, b(\mathbf{X}_i^T \beta + \tilde{f}_n(t_i))) - \operatorname{Pen}(f)$. The expression of $b(\mathbf{X}_i^T \beta + \tilde{f}_n(t_i))$ can indeed be simplified using first order conditions of step 1. Due to the non-linearity of our procedure, we choose here to keep a loglikelihood approach.
3. $\hat{f}_n = \tilde{f}_{\hat{\beta}_n}$.

Note also that an usual estimation procedure used in generalized models is quasi-likelihood maximization. For details, we refer among others to Severini & Staniswalis (1994) or McCullagh & Nelder (1989). In GPLM, quasi-likelihood estimation was developed for example by Chen (1987) and Mammen & Van der Geer (1997).

2.2 Discrete wavelet transform

The aim of the present work is to introduce a wavelet penalty in estimation. The idea is to use the wavelet representation of the functional part we wish to estimate through this penalty. For more precision on wavelets, the reader is referred to Daubechies (1992), Y. Meyer (1992) or Mallat (1999).

Let $(L^2[0,1], \langle \cdot, \cdot \rangle)$ be the space of squared-integrable functions on $[0,1]$ endowed with the inner product $\langle f, g \rangle = \int_{[0,1]} f(t)g(t) dt$. Throughout the paper we assume that we are working within an r -regular ($r \geq 0$) multiresolution analysis of $(L^2[0,1], \langle \cdot, \cdot \rangle)$, associated with an orthonormal basis generated by dilatations and translations of a compactly supported scaling function, $\varphi(t)$, and a compactly supported mother wavelet, $\psi(t)$. For simplicity reasons, we will consider periodic wavelet bases on $[0,1]$.

For any $j \geq 0$ and $k = 0, 1, \dots, 2^j - 1$, let us define $\varphi_{j,k}(t) = 2^{j/2}\varphi(2^j t - k)$ and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$. Then for any given resolution level $j_0 \geq 0$ the family

$$\left\{ \varphi_{j_0,k}, k = 0, 1, \dots, 2^{j_0} - 1; \psi_{j,k}, j \geq j_0; k = 0, 1, \dots, 2^j - 1 \right\}$$

is an orthonormal basis of $L^2[0,1]$. Let f be a function of $L^2[0,1]$; if we denote by $c_{j_0,k} = \langle f, \varphi_{j_0,k} \rangle$ ($k = 0, 1, \dots, 2^{j_0} - 1$) the scaling coefficients and by $d_{j,k} = \langle f, \psi_{j,k} \rangle$ ($j \geq j_0; k = 0, 1, \dots, 2^j - 1$) the wavelet coefficients of f , the function f can then be decomposed as follows:

$$f(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k} \varphi_{j_0,k}(t) + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} d_{j,k} \psi_{j,k}(t), \quad t \in [0,1].$$

Yet in practice, we are more concerned with discrete observation samples rather than continuous. Consequently we are more interested by the discrete wavelet transform (DWT). Given a vector of

real values $\mathbf{e} = (e_1, \dots, e_n)^T$, the discrete wavelet transform of \mathbf{e} is given by $\boldsymbol{\theta} = \Psi_{n \times n} \mathbf{e}$, where $\boldsymbol{\theta}$ is an $n \times 1$ vector comprising both discrete scaling coefficients, $\theta_{j_0, k}^S$, and discrete wavelet coefficients, $\theta_{j, k}^W$. The matrix $\Psi_{n \times n}$ is an orthogonal $n \times n$ matrix associated with the orthonormal periodic wavelet basis chosen, where one can distinguish the *Blocs* spanned respectively by the scaling functions and the wavelets.

Note that if \mathbf{F} is a vector of function values $\mathbf{F} = (f(t_1), \dots, f(t_n))^T$ at equally spaced points t_i , then the corresponding empirical coefficients $\theta_{j_0, k}^S$ and $\theta_{j, k}^W$ are related to their continuous counterparts $c_{j_0, k}$ and $d_{j, k}$ with a factor $n^{-1/2}$. It is worthy to remark also that because of orthogonality of $\Psi_{n \times n}$, the inverse DWT is simply given by $\mathbf{F} = \Psi_{n \times n}^T \boldsymbol{\theta}$. If $n = 2^J$ for some positive integer J , Mallat (1989) propose a fast algorithm, that requires only order n operations, to compute the DWT and the inverse DWT.

2.3 Assumptions and asymptotic minimaxity

Let $\|\cdot\|$ denotes the euclidean norm on \mathbb{R}^p and $\|h\|_n^2 = \frac{1}{n} \sum_{i=1}^n h(t_i)^2$ for any function h . To ameliorate the comprehension of the results and the assumptions, the subscript 0 will identify in the following the true values of the model.

Due to the use of the Discrete Wavelet Transform described above, we will consider in the following that the functional part is observed on an equidistant sample $t_i = \frac{i}{n}$, and that the sample size satisfy $n = 2^J$ for some positive integer J .

We first introduce assumption (A1) which ensures the identifiability of the model:

(A1) $\frac{1}{n} X^T X$ converges to a strictly positive matrix when n goes to infinity, and $\frac{1}{n} X^T F_0$ goes to 0 when n goes to infinity, with $F_0 = (f_0(t_1), \dots, f_0(t_n))^T$.

Define $H = X(X^T X)^{-1} X^T$ the projection matrix on the space generated by the columns of X . The matrix H admits a rank and thus a trace equal to p . If $h_i = \mathbf{X}_i (X^T X)^{-1} \mathbf{X}_i^T$ denotes the i^{th} diagonal term of H , this means that $\sum h_i = p$. We moreover suppose that:

(A2) $h = \max_{i=1 \dots n} h_i \rightarrow 0$

This assumption is very usual (see *e.g.* Huber (1981)).

Concerning the loglikelihood function, we assume that:

(A3) $\sup_{\|\eta - \eta_0\|_n \leq 2(C_\infty + \|f\|_\infty)} \sup_i \ddot{b}(\eta_i) \leq \ddot{b}_\infty < \infty$. We also suppose that $\min_i \ddot{b}(\mathbf{X}_i^T \beta_0 + f_0(t_i)) \geq \ddot{b}_0 > 0$ and $\max_i \ddot{b}(\mathbf{X}_i^T \beta_0 + f_0(t_i)) \leq \ddot{b}_\infty < \infty$. Moreover, $\frac{1}{n} \sum_{i=1}^n \ddot{b}(\mathbf{X}_i^T \beta_0 + f_0(t_i)) \mathbf{X}_i^T \mathbf{X}_i$ converges to a positive matrix Σ_0 .

Recall that the function $\ddot{b}(\cdot)$ is associated to the variance of the observations and that one has $\ddot{b} > 0$.

Then some more restrictive assumptions are made on the form of the distribution:

(A4.1) There exist a constant $a > 0$ such that $\max_{i=1, \dots, n} \mathbb{E} \left[\exp(\dot{\ell}(Y_i, \mathbf{X}_i^T \beta + f(t_i))^2 / a) \right] \leq a$,

(A4.2) There exist constants $K, \sigma_0^2 > 0$ such that

$$\max_{i=1, \dots, n} K^2 \left(\mathbb{E}[\exp \left(\left| \dot{\ell}(Y_i, \mathbf{X}_i^T \beta + f(t_i)) \right| / K^2 \right)] - 1 \right) \leq \sigma_0^2.$$

Assumption (A4.1) corresponds to exponential tails and is weaker than assumption (A4.2), which corresponds to sub-Gaussian tails.

We aim to control the correlation between the linear and the nonparametric parts of the model. Following Rice (1986) or Speckman (1988) we decompose the components of the design matrix X into a sum of a deterministic function of $L^2[0, 1]$ and a noise term. More precisely, the (i, j) -component of X , say $x_{i,j}$, is supposed to take the form $x_{i,j} = g_i(t_j) + \xi_{i,j}$ with functions $(g_i)_{i=1, \dots, p}$ forming an orthogonal family on $(L^2[0, 1], \langle \cdot, \cdot \rangle)$ and with $\xi_{i,j}$ denoting a realization of a random variable ξ_i . The variables $(\xi_i)_{i=1, \dots, n}$ are supposed to be centered, independent, with finite variance. We can easily see that the orthogonality of the family $(g_i)_{i=1, \dots, p}$ ensures that the matrix $\frac{1}{n} X^T X$ goes to a strictly positive matrix when n goes to infinity. If in addition the system $(g_i)_{i=1, \dots, p}$ satisfies $\int_{[0,1]} f(t) g_i(t) dt = 0$, then assumption (A1) and consequently identifiability are guaranteed.

We also make an assumption on the distribution of the random variables ξ_i , and of course we suppose we control the regularity of the functions g_i .

(A_{corr}) $\forall j = 1, \dots, p, i = 1, \dots, n, X_{i,j} = g_j(t_i) + \xi_{i,j}$, with polynomial functions g_j of degree less or equal than the number of vanishing moments of the wavelet considered. For all $j = 1, \dots, p$, $(\xi_{i,j})_{i=1, \dots, n}$ is a n -sample such that $\max_{i=1, \dots, n} \mathbb{E} \left[\exp(\xi_{i,j}^2/a_j) \right] \leq a_j$, for given constants $a_j > 0$.

2.3.1 Nonadaptive case

Assume $J(\cdot)$ is a given criterion on the functions from $[0, 1]$ to the positive real line. We introduce the function class $\mathcal{A} = \{g, J(g) \leq C\}$. Let us recall the definition of the entropy of a subspace:

Definition 1. Let \mathcal{F} be a subset of a metric space (\mathcal{L}, d) of real-valued functions. The δ -covering number $N(\delta, \mathcal{F}, d)$ of the space \mathcal{F} is the smallest number N such that there exist a_1, \dots, a_N such that for each $a \in \mathcal{F}$ one have $d(a, a_i) \leq \delta$ for some $i \in \{1, \dots, N\}$. The δ -entropy $\mathcal{H}(\delta, \mathcal{F}, d)$ of the space \mathcal{F} is defined as $\mathcal{H}(\delta, \mathcal{F}, d) = \log N(\delta, \mathcal{F}, d)$.

We here suppose that the δ -entropies of the subspace \mathcal{A} for the distance associated to the norm $\|\cdot\|_n$ behave like

$$\limsup_{n \rightarrow \infty} \sup_{\delta > 0} \delta^\nu \mathcal{H}(\delta, \mathcal{A}, \|\cdot\|_n) < \infty,$$

for a given $0 < \nu < 2$.

The penalty in equation (5) is chosen according to the two assumptions (A6) and (A7):

(A5) For any function h , $\frac{\lambda v_n^2}{n} \text{Pen}(h) \geq J(h)$ with $v_n = n^{1/(2+\nu)}$.

(A6) $f \mapsto K_n(f, \beta) = \sum_{i=1}^n \ell(y_i, \mathbf{X}_i^T \beta + f(t_i)) - \lambda \text{Pen}(f)$ is concave.

When the special structure given in (A_{corr}) is assumed, we are willing to exploit it through a penalty on wavelet coefficients :

(A7) The penalty $\text{Pen}(h)$ applies only to the wavelet decomposition coefficients $(\theta_{j,k}^W)_{j \geq j_S, k \in \mathbb{Z}}$ of the function h .

We are now in position to give our first asymptotic result.

Theorem 2. Suppose assumptions (A1) to (A3), (A4.1), (A5) and (A6) hold and $J(f_0) < \infty$. Let β be a given p -vector such that $\sqrt{n}\|\beta - \beta_0\| \leq c$. Define $\tilde{f}_\beta = \underset{\{f, \|f\|_\infty \leq C_\infty\}, \beta \in \mathbb{R}^p}{\operatorname{argmax}} K_n(f, \beta)$. Then,

$$\begin{aligned} v_n \|\tilde{f}_\beta - f_0\|_n &= \mathcal{O}_{\mathbb{P}}(1) \\ J(\tilde{f}_\beta) &= \mathcal{O}_{\mathbb{P}}(1). \end{aligned}$$

Define $\hat{\beta}_n = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}} K_n(\tilde{f}_\beta, \beta)$. Then,

$$v_n \|\hat{\beta}_n - \beta_0\| = \mathcal{O}_{\mathbb{P}}(1),$$

If in addition the covariates of the linear part admit a representation of the form given in assumption (A_{corr}) and if the penalty satisfies (A7), then:

$$\sqrt{n} \|\hat{\beta}_n - \beta_0\| = \mathcal{O}_{\mathbb{P}}(1).$$

The results still hold if the number p of regression covariates goes to infinity provided the sequence $n^{(v-2)/(v+2)}p$ goes to 0 when n goes to infinity and the sequence $n^{-v/(2+v)}p$ is bounded.

The proof is given in Appendix. The main keys are controls given by Van der Geer (2000), relying on the entropy.

Note that minimax optimality is obtained both for the linear predictor and the nonparametric estimator. The condition of correlation (A_{corr}) under which the optimality is attained is similar to the one given in Gannaz (2007a). This condition on design covariates appears to be more flexible than Rice's (1986) or Speckman's (1988) in the sense that the maximal degree of the polynomial functions intervening in the covariates depends on the number of vanishing moments of the wavelet, instead of depending on the regularity of the function f . Compared to Mammen & Van der Geer (1997), the assumptions seem much more weaker. Note that without correlation conditions both estimators attain nonparametric convergence rates.

The fact that the results hold for a number of covariate p going to infinity allows to have many covariates in the linear part. This remark can be useful for dimension reduction modelling, where

the number of covariates is large. However the rate of convergence for p may be poor when the regularity of the function f is poor.

In order to illustrate the general framework in which we gave the asymptotic behaviour, let us consider the penalty proposed in Antoniadis (1996). To exploit the sparsity of wavelet representations, we will assume that f belongs to a Besov space on the unit interval, $\mathcal{B}_{\pi,r}^s([0,1])$, with $s + 1/\pi - 1/2 > 0$. The last condition ensures in particular that an evaluation of f at a given point makes sense. For a detailed overview of Besov spaces we refer to Donoho & Johnstone (1998).

Corollary 1. *Suppose f belongs to a Besov ball $\mathcal{B}_{\pi,r}^s(R)$ with $s > 1/2$, $0 < s - 1/2 + 1/\pi$, $\pi > 2/(1 + 2s)$ and $1/\pi < s < \min(R, N)$, where N denotes the number of vanishing moments of the wavelet ψ . Take the penalty: $\text{Pen}(f) = \sum_{j=j_s}^n 2^{2js} \sum_k |\theta_{j,k}^W|^2$ where $\theta_{j,k}^W$ are the wavelet coefficients of f . Assume conditions of Theorem 2, (A7) and (A_{corr}) hold.*

If $\lambda \sim n^{-2s/(1+2s)}$, we can deduce from Theorem 2 that

$$\begin{aligned} \|\widehat{f}_n\|_{s,2,\infty} &= \mathcal{O}_{\mathbb{P}}(1), \\ n^{s/(1+2s)} \|\widehat{f}_n - f_0\|_n &= \mathcal{O}_{\mathbb{P}}(1) \\ \sqrt{n} \|\widehat{\beta}_n - \beta_0\| &= \mathcal{O}_{\mathbb{P}}(1). \end{aligned}$$

where $\|f\|_{s,2,\infty} = \sup_{j \geq j_s} 2^{j(s-1/2)} \left(\sum_k |\theta_{j,k}^W|^2 \right)^{1/2}$.

Proof. Birgé & Massart (2000) establish the entropy of Besov balls $\mathcal{B}_{\pi,\infty}^s(1)$, with $2/(1 + 2s) < \pi$, is $\nu = 1/s$. One can see that $\frac{\lambda v_n^2}{n} \text{Pen}(f) \sim \|f\|_{s,2,\infty}^2 = J(f)$. Consequently the δ -entropy of the functional set $\{f, J(f) \leq c\}$ can be bounded up to a constant by $\delta^{-1/s}$. We thus can apply Theorem 2. \square

One drawback of this estimation procedure is its non adaptivity: the optimal value of the smoothing parameter λ depends on the regularity s of the function, which is unknown. Actually the way the penalty term is used cannot lead to adaptivity since it is closely linked to the norm of the functional space where the function f lies. Another penalty type may be introduced.

2.3.2 Adaptive case

This section deals with the introduction of an adaptive penalty. We choose to use a ℓ^1 -penalty on the wavelet coefficients. It is well known that such a penalty leads to soft-thresholding wavelet estimators, which are adaptive. We refer to Antoniadis & Fan (2001) for general theory on penalization on wavelets coefficients and to Loubes & Van der Geer (2002) for the use of the ℓ^1 penalty in functional models.

Our asymptotic results for ℓ^1 penalty is the following:

Theorem 3. *Suppose assumptions (A0) to (A3) and assumptions (A4.2) and (A6) hold. Let β be a given p -vector such that $\sqrt{n}\|\beta - \beta_0\| \leq c$. Let K_n be given by equation (5), with the penalty: $\text{Pen}(f) = \lambda \sum_{i=1}^n |\theta_i^W|$ where (θ_i^W) are the wavelet coefficients of f . Define $\tilde{f}_\beta = \underset{\{f, \|f\|_\infty \leq C_\infty\}, \beta \in \mathbb{R}^p}{\operatorname{argmax}} K_n(f, \beta)$. Then, for $\lambda \sim \sqrt{\log(n)}$, we have*

$$\left(\frac{n}{\log(n)}\right)^{s/(1+2s)} \|\tilde{f}_\beta - f_0\|_n = \mathcal{O}_{\mathbb{P}}(1).$$

Define $\hat{\beta}_n = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}} K_n(\tilde{f}_\beta, \beta)$. Then,

$$v_n \|\hat{\beta}_n - \beta_0\| = \mathcal{O}_{\mathbb{P}}(1),$$

If in addition the covariates of the linear part admit a representation of the form given in assumption (A_{corr}), then:

$$\sqrt{n} \|\hat{\beta}_n - \beta_0\| = \mathcal{O}_{\mathbb{P}}(1).$$

The results still hold if the number p of regression covariates goes to infinity provided the sequence $\left(\frac{n}{\log(n)}\right)^{(2s-1)/(1+2s)} p$ goes to 0 when n goes to infinity and the sequence $\left(\frac{n}{\log(n)}\right)^{1/(1+2s)} p$ is bounded.

The proof is given in Appendix and relies on M-estimation techniques of Van der Geer (2000).

As noted before, assumption (A4.2) is more restrictive than assumption (A4.1). The results of the Theorem could probably be extended to exponential tails distributions, weakening assumption (A4.2). We refer to discussion on page 134 of Van der Geer (2000) (Corollaries 8.3 and 8.8) for a discussion on the price to pay to release assumption (A4.2).

The adaptivity is acquired. As explained before, it gives the possibility of a computable procedure, without need of a cross-validation step. Yet, the parameter λ is only chosen among an asymptotic condition and a finite sample application arises that the exact choice of this parameter is important. This will be discussed hereafter. Note also that the link with soft-thresholding is not evident, but appears in the iterative implementation of the estimators in next section.

The optimality of the functional part estimation is of course still available in a generalized functional model. In such models Antoniadis et al. (2001) and Antoniadis & Sapatinas (2001) propose an adaptive estimation when the variance is respectively cubic or quadratic. These assumption have to be compared with (A3) and (A4.2). Recently, Brown et al. (2008) introduced a method which consists in a transformation on the observations, based on the central limit theorem, in order to be able to use Gaussian framework's results. Yet, even if the asymptotic results are satisfactory, the implementation needs an important number of observations. We will see in numerical study that our procedure is easier to compute. Note also that it is available with the presence of the linear part.

2.4 Algorithm

This section is only devoted to the adaptive ℓ^1 -type penalty. Similar algorithms are available for other penalties but a cross-validation procedure should be elaborated because of the lack of adaptivity. It is worthy to see that the proposed estimators can be easily computed, by the way of iterative algorithms. A short simulation study is also given to evaluate the performance of the estimation with finite samples.

The paper of Müller (2001) concludes that a loglikelihood maximisation is preferable to a *profile likelihood* estimation (we recall this procedure consists in the maximisation of a modified loglikelihood criterion when estimating the parametric part). According to the author, *backfitting* computation gives in general better numerical values in estimation than others methods in GPLM. In a Gaussian framework, the same conclusion was observed in Gannaz (2007a) (note that time differences with Gannaz (2007b) are due to an improvement of the *backfitting* algorithm used). We

therefore implement a *backfitting* algorithm.

Backfitting: Let $\beta^{(0)}$ be a given p -dimensional vector. For each iteration k , do:

Step 1 $f^{(k+1)} = \underset{f}{\operatorname{argmax}} K(f, \beta^{(k)})$

Step 2 $\beta^{(k+1)} = \underset{\beta}{\operatorname{argmax}} K(f^{(k+1)}, \beta)$

The algorithm is stopped either when a maximal number of iterations κ is attained or when the algorithm is stabilized, *i.e.* when $\|\beta^{(k)} - \beta^{(k-1)}\| \leq \delta \|\beta^{(k-1)}\|$ for a given tolerance value δ . The returned values are $\hat{\beta}_n = \beta^{(K)}$ and $\hat{f}_n = f^{(K)}$ with K maximal number of iterations of the algorithm.

To compute each of the two steps we apply a classical Fisher-scoring algorithm, detailed among others page 40 of McCullagh & Nelder (1989). Usual in generalized models, this algorithm consists in building new variables of interest by a gradient descending method, in order to apply a ponderate regression on these new variables. Recall the notations of the GPLM given in equation (4): one have $\eta(X, t) = X^T \beta + f(t)$ and $\mu(X, t) = \dot{b}(\eta(X, t))$. We will omit the dependence to (X, t) for the sake of simplicity.

Step 1:

Note $f^{(k,0)} = f^{(k)}$. Repeat the following iteration for $j = 0 \dots J_1 - 1$:

Let $\eta^{(k,j)} = X\beta^{(k)} + f^{(k,j)}$.

We introduce $Y^{(k,j)} = f^{(k,j)} + (y - \mu^{(k,j)}) \frac{d\eta}{d\mu} \Big|_{\mu=\mu^{(k,j)}}$ and $W^{(k,j)} = \operatorname{diag} \left(\frac{d\eta}{d\mu} \Big|_{\mu=\mu^{(k,j)}} \right)$.

With an ℓ^1 -penalty, we establish that $f^{(k,j+1)}$ is a nonlinear wavelet estimator for the observations $Y^{(k,j)}$, obtained by soft-thresholding of the wavelet coefficients. The threshold levels are $\lambda \Psi W^{(k,j)-1} \Psi^T \mathbf{1}_{n \times 1}$ where Ψ denotes the forward wavelet transform and Ψ^T the inverse wavelet transform.

Take $f^{(k+1)} = f^{(k, J_1)}$.

Step 2:

Define $\beta^{(k,0)} = \beta^{(k)}$. Repeat the following iteration for $j = 0 \dots J_2 - 1$:

Let $\tilde{\eta}^{(k,j)} = X\beta^{(k,j)} + f^{(k+1)}$.

We introduce $\tilde{Y}^{(k,j)} = X\beta^{(k)} + (y - \tilde{\mu}^{(k,j)}) \frac{d\eta}{d\mu} \Big|_{\mu=\tilde{\mu}^{(k,j)}}$ and $\tilde{W}^{(k,j)} = \left(\frac{d\eta}{d\mu} \Big|_{\mu=\tilde{\mu}^{(k,j)}} \right)$.

Then $\beta^{(k,j+1)}$ is the regression parameter of $Y^{(k,j)}$ on X with ponderations $\tilde{W}^{(k,j)}$, i.e. $\beta^{(k,j+1)} = (X^T \tilde{W}^{(k,j)-1} X)^{-1} X^T \tilde{W}^{(k,j)-1} \tilde{Y}^{(k,j)}$.

Take $\beta^{(k+1)} = \beta^{(k,J_2)}$.

Maximal number of iterations J_1 and J_2 can be fixed to 1 to simplify the algorithm (this is what is proposed actually in Müller (2001)). In computation studies, no main difference is observed while modifying parameters J_1 and J_2 . We therefore also decide to fix them equal to 1.

The initialization values are set as follows: for all $i = 1, \dots, n$, $f^{(0)}(t_i) = G(y_i)$, with G the link function associated to the model (with a slight modification if the value does not exist) and for all $j = 1, \dots, p$, $\beta_j^{(0)} = 0$.

In the particular Gaussian case, the two steps are non iterative. We explicit how the estimators are implemented in a Gaussian framework to ameliorate the comprehension of the algorithm:

Step 1: The iterate $f^{(k+1)}$ is the wavelet estimator for the observations $y - X^T \beta^{(k)}$, with a thresholding on wavelet coefficients with an uniform threshold level λ .

Step 2: We obtain $\beta^{(k+1)}$ by a maximum likelihood estimation on $y - f^{(k+1)}$; this means that one have $\beta^{(k+1)} = (X^T X)^{-1} X^T (Y - f^{(k+1)})$.

We can recognize the *backfitting* algorithm studied in Chang & Qu (2004), Fadili & Bullmore (2005) and Gannaz (2007b). In a Gaussian framework, the variance in observations is constant and consequently the threshold level is uniform. In a generalized framework, the matrix W ponderates the threshold level to take into account the inhomogeneity of the variance.

In generalized functional models, i.e. without the presence of a linear part, the numerical implementation of the estimators proposed here has already been explored by Sardy et al. (2004). The authors propose an interior point algorithm based on the dual maximisation problem. Comparing to this resolution scheme, our procedure has the advantage of an easy interpretation of the different steps in the algorithm. As noted previously, wavelet estimators have also been explored by Antoniadis et al. (2001), Antoniadis & Sapatinas (2001) and Brown et al. (2008). These

papers need to aggregate the data into a given number of bins. If the two first cited papers allow to choose small size of bins, it appears to be quite a constraint for the third one. Actually, it is worthy noticing the simplicity of our algorithm, *e.g.* comparing with those implementations.

2.5 Simulations

In this subsection, we give some simulation results. All the calculations were carried out in MATLAB 7.6 on a Unix environment. For the DWT, we used the WaveLab toolbox developed by Donoho and his collaborators at the Statistics Department of Stanford University (see <http://www-stat.stanford.edu/~wavelab>).

We simulate $n = 2^8$ observations. The covariates are written according to assumption (A_{corr}): $x_i = g(i/n) + \xi_i$, with ξ_i independent and identically distributed variables following a standard normal distribution, and with $g(x) = 30(x - 0.5)^4 - 6(x - 0.5)^2 + (x - 0.5)$. Three functional parts f were considered : we will refer at

- the Sinus function, for a smooth function, combination of sinusoidal functions,
- the *Blocs* function for a piecewise constant function,
- and the *Pics* function for a function presenting high variations.

These functions are given in Figure 1.

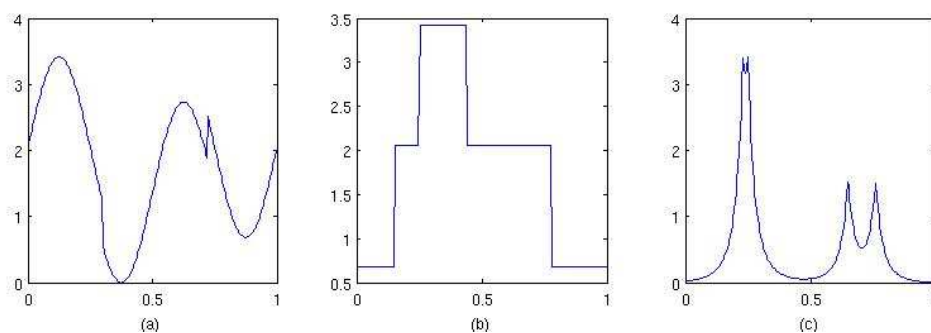


Figure 1: Functional part of the generalized partially linear model. Figure (a) corresponds to the *Sinus* function, Figure (b) to the *Blocs* function and Figure (c) to the *Pics* function.

We will more precisely study the estimation quality for

- a Gaussian distribution; observations are $y_i \sim \mathcal{N}(\eta_i, \sigma^2)$, with $\sigma^2 = \phi$,
- a Binomial distribution; observations are y_i such that $y_i \times m \sim \mathcal{B}(\mu_i, m)$ with $m = \phi^{-1}$ and the logit link function $\mu_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$,
- a Poisson distribution; observations are $y_i \sim \mathcal{P}(\mu_i)$, with the link function $\mu_i = \exp(\eta_i)$,

as these distributions seem to be the most frequently encountered in modelisation.

To conclude with respect to the quality of estimation, we will give the mean value estimated for the parameter β as well as the standard deviation of the estimation. For the nonparametric part, we will evaluate the MISE, which is the empirical estimate for $\mathbb{E} \left[\int_{[0,1]} \left(\hat{f}_n(t) - f_0(t) \right)^2 dt \right]^{1/2}$. All results were obtained on 500 simulations with the same covariates x_i and the same functional parts f . A maximal number of $\kappa = 5000$ iterations was taken and the tolerance value defined above was equal to $\delta = 10^{-20}$ when applying the algorithm.

2.5.1 Preliminary study : choice of the threshold level

The asymptotic behaviour of the estimators only defines the threshold level λ up to a constant. Yet, numerical implementation needs to determine the exact threshold level λ in the algorithm. In practice, one can see it has an important impact on the quality of estimation. Figure 2 gives the evolution of the MISE with respect to the threshold level in the GPLM with different distributions. One can observe that the threshold levels attaining the minima are very different among the distribution.

With a Gaussian distribution, following Donoho et al. (1995), we choose $\lambda = \sqrt{2\phi \log(n)}$. This choice overevaluates the optimal threshold level in many cases but is the most often encountered in practice.

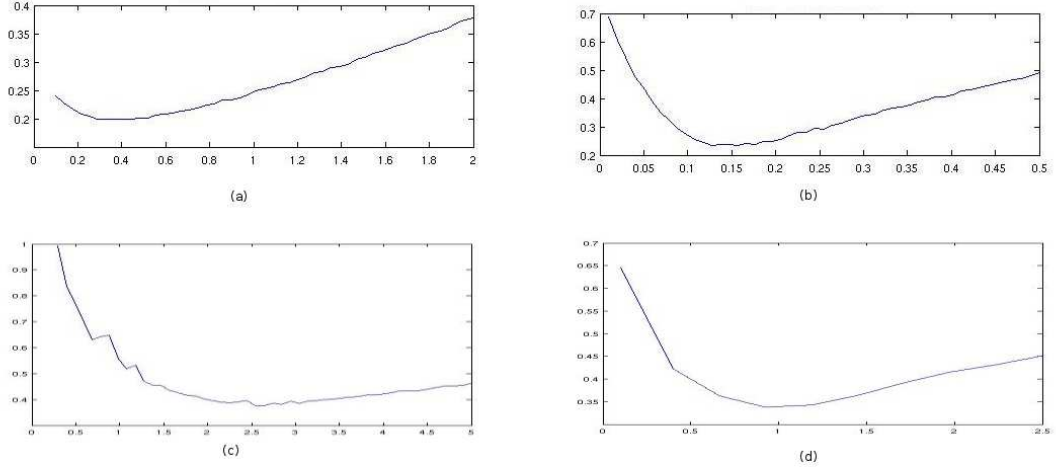


Figure 2: Evolution of the MISE with respect to the threshold level in a GPLM with the *Sinus* function. Figure (a) corresponds to a Gaussian distribution, Figure (b) to a binomial distribution, Figure (c) to a Poisson distribution and Figure (d) to an Exponential distribution.

Binomial distribution

Due to homogeneity reasons, it seems well-adapted to fix a threshold level of the form $\lambda = c\sqrt{\phi \log(n)}$, where ϕ corresponds to the dispersion parameter in distribution (1) and c denotes a positive constant.

To ensure this conjecture in a Binomial setup, we plot the optimal threshold obtained for different value of the Binomial parameter m . The study was done on a generalized functional model (GFM) and on a GPLM. Figure 3 confirms that the chosen form seems appropriated.

Linear fittings are given in Table 1. The linear fitting is coherent provided the R^2 coefficients. The constant c obtained by linear fitting is varying with respect to the simulated samples, but the observed variations are small. The mean value is 0.4997 and the standard deviation is 0.011. The conjecture of a uniform constant seems acceptable. We therefore choose to take the value $c = 0.5$ in the following for Binomial distributions. Note that due to the central limit theorem, one would have expected to take $c = \sqrt{2}$ for large values of the parameter m , but our simulation study show that this would lead to an oversmoothing for small values of m .

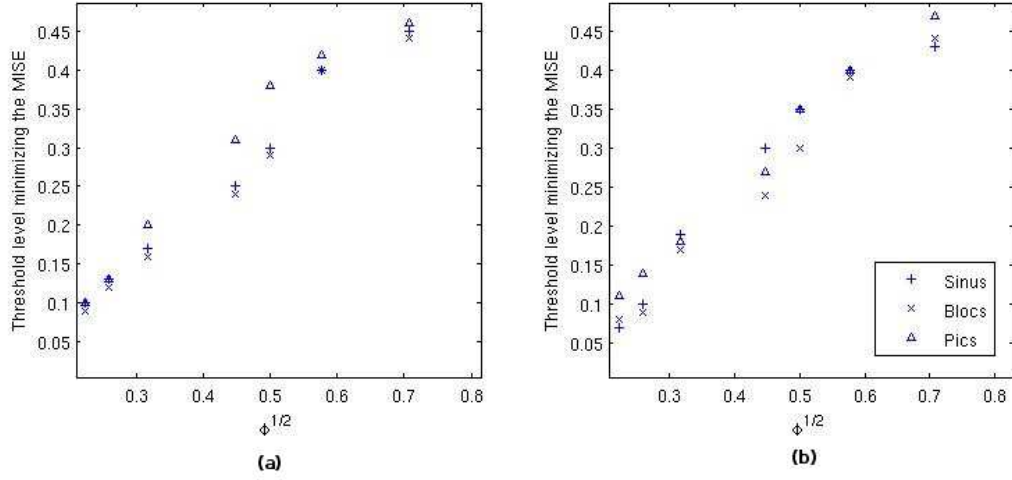


Figure 3: Evolution of the threshold minimizing the MISE when estimating the *Sinus*, *Blocs* and *Pics* functions in a Binomial setup, with respect to $\sqrt{\phi}$. In a GFM on Figure (a) and in a GPLM in Figure(b). Calculations were done on 100 samples of size $n = 2^8$.

Generalized functional model			
Function	<i>Sinus</i>	<i>Blocs</i>	<i>Pics</i>
R^2 coefficient	0.984	0.977	0.961
constant c	0.483	0.488	0.512

Generalized partially linear model			
Function	<i>Sinus</i>	<i>Blocs</i>	<i>Pics</i>
R^2 coefficient	0.947	0.981	0.989
constant c	0.510	0.507	0.498

Table 1: Numerical indexes for the regression of the threshold level that minimizes the MISE with respect to $\sqrt{\phi \log(n)}$ with a Binomial distribution. Calculations were done on 100 samples of size $n = 2^8$.

Note all the calculations were done with a fixed sample size $n = 2^8$. To better evaluate the form of the optimal threshold in practice, one could also study the evolution of the threshold level with respect to the sample size n .

Poisson distribution

Estimation in a Poisson functional model has been more intensively explored. Note that Sardy et al. (2004) propose a threshold level. The main drawback is that the level given depends on the estimated function. Yet, the choice is based on an universal large deviation inequality which does not seem to be well-adapted in this procedure. Indeed, due to the iterative interpretation of the estimation, the inhomogeneity of the variance of the observations is taken into account within the estimation.

Recently, Reynaud-Bouret & Rivoirard (2010) have developed a procedure based on wavelet hard-thresholding estimation for Poisson regression. In their estimation the thresholding step is defined directly and not through a penalization procedure like here. The authors then present a detailed numerical study showing the high instability of the optimal threshold level with respect to the estimated function.

In our procedure, we can hope for a better stability of the threshold level, considering the fact that it takes into account the variance of the pseudo-wavelet coefficients at each iteration.

Figure 4 represents the evolution of the threshold level which minimizes the MISE with respect to $\sqrt{\log(n)}$. Table 2 gives some numerical results associated with Figure 4.

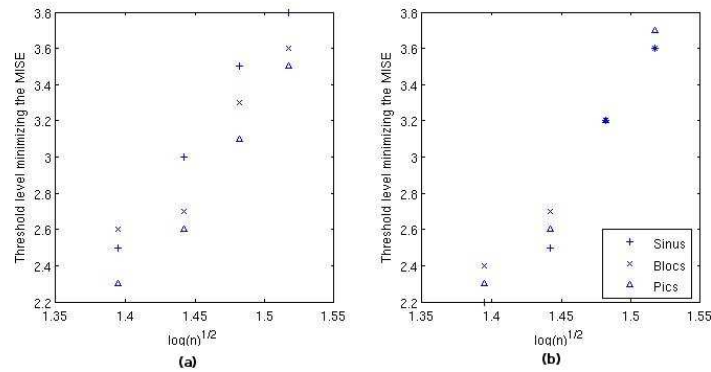


Figure 4: Evolution of the optimal threshold with respect to $\sqrt{\log(n)}$ with a Poisson distribution. In a GFM on Figure (a) and in a GPLM in Figure(b). Calculations were done on 100 samples of size $n = 2^8$.

Generalized functional model				
Sample size n	7	8	9	10
Mean value	1.699	1.783	2.005	2.094
Standard deviation	0.086	0.110	0.099	0.072

Generalized partially linear model				
Sample size n	7	8	9	10
Mean value	1.584	1.675	1.944	2.094
Standard deviation	0.056	0.053	0.028	0.027

Table 2: Mean value and standard deviation of the ratio between the threshold level that minimizes the MISE and $\sqrt{\log(n)}$ with a Poisson distribution. Calculations were done on 100 samples for each function *Sinus*, *Blocs* and *Pics*.

We observe that the optimal threshold level does not vary significantly with respect to the function estimated. This is an advantage on the estimation developed by Reynaud-Bouret & Rivoirard (2010). This also confirms that the threshold level proposed by Sardy et al. (2004) is not adapted here. It seems in fact that the threshold level advised by Sardy et al. (2004) is more convenient for a uniform procedure like in Reynaud-Bouret & Rivoirard (2010) than for the iterative estimation scheme implemented here. Moreover it appears that the presence of a linear part in the model does not change the threshold level. This stability of the estimation procedure is a interesting property. Recall it can be explained by the evaluation of the variance of the pseudo-variables in the iterative algorithm.

Note that we do not obtain an optimal threshold level proportional to $\sqrt{\log(n)}$. As the theoretical result is asymptotic, we should study larger values of the sample size n . According to the results in Table 2, one may choose a constant approximatively equal to 2 in a functional model and in a GPLM. We therefore choose to take the value $c = 2$ in the following for Poisson distribution.

Remark: In a Gaussian or a Binomial regression, one may need the dispersion parameter ϕ . Actually in literature, it is classically estimated at each iteration by

$$\phi^{(k)} = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \mu_i^{(k)})^2}{\dot{b}(\eta^{(k)})}.$$

Due to the bad quality of this estimator in GPLM, we prefer to consider in this paper that the dispersion parameter is known. In a Gaussian model, Gannaz (2007b) proposed an efficient QR-based estimator for ϕ . It would be worthy to explore whether it could be extended to generalized models.

2.5.2 Example 1: Gaussian distribution

Example 1 deals with a Gaussian model. The Gaussian distribution implementation is not a novelty for this estimation procedure, and we refer to Chang & Qu (2004), Fadili & Bullmore (2005) and Gannaz (2007b) for detailed studies on simulated or real values data. We briefly consider this case in order to have a comparison base for other distributions.

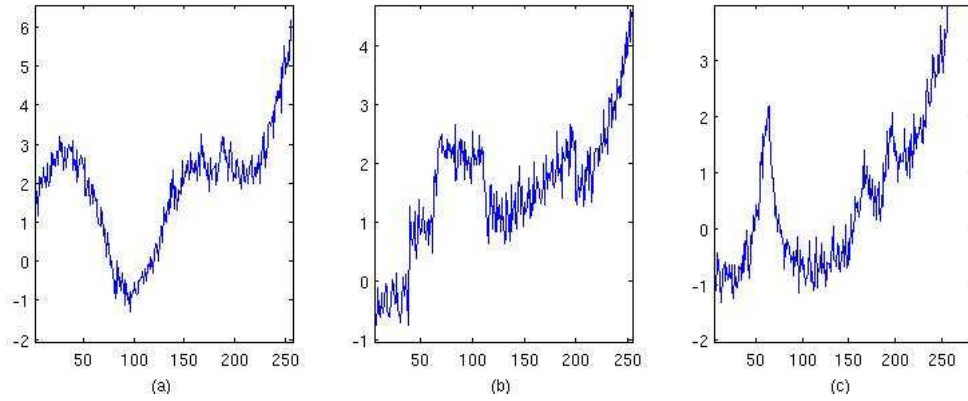


Figure 5: An example of a simulated data set in Example 1, with the function *Sinus* in Figure (a), the function *Blocs* in Figure (b) and the function *Pics* in Figure (c).

The signal-to-noise-ratio (SNR) of a signal is defined as the norm of the ratio of the mean value with respect to the standard deviation. In GPLM the SNR for the nonparametric part, noted SNR_f

and the SNR for the linear part of the model, noted SNR_{β} , are respectively equal to

$$SNR_f^2 = \frac{1}{n} \sum_{i=1}^n \frac{f_0(t_i)^2}{\phi \ddot{b}(\mathbf{x}_i^T \boldsymbol{\beta}_0 + f_0(t_i))},$$

$$\text{and } SNR_{\beta}^2 = \frac{1}{n} \sum_{i=1}^n \frac{(\mathbf{x}_i^T \boldsymbol{\beta}_0)^2}{\phi \ddot{b}(\mathbf{x}_i^T \boldsymbol{\beta}_0 + f_0(t_i))}.$$

With a high SNR, say approximatively 9, one can expect a good quality of estimation, while with a small value, like 2, the quality of estimation cannot be satisfying.

Table 3 gives quality indexes for the estimation of the model on 500 simulations. As already noted in the different papers using this estimation procedure, the numerical results appear to be of fairly good quality.

Function	Estimation of β		Estimation of f		Time
	Mean SNR	Mean estimation	Mean SNR	MISE	
<i>Sinus</i>	8.7	0.9993(0.0098)	9.1	0.1639	1.31(326)
<i>Blocs</i>	6.3	0.9991(0.0168)	9.0	0.2197	0.49(118)
<i>Pics</i>	2.2	1.0002(0.0296)	1.6	0.3367	0.64(153)

Table 3: Measures of quality the estimates over the 500 simulations in Example 1 with $n = 2^8$. For the parameter β , the true value is 1 and the value given is the mean value and standard deviation appears in brackets. In the column Time, the mean of numbers of iterations in the algorithm is given in brackets.

For the value of the signal-to-noise ratio ($SNR_f = 9$) adopted in our simulations for the nonparametric part, the estimator is nearly able to detect the discontinuity of the *Sinus* function, as shown in Figure 6 (a). Results for the nonparametric part are very similar to those obtained in a nonparametric signal denoising, without a linear part. The asymptotic behaviour of the estimates effectively states that the presence of the linear part does not affect the estimation of the nonparametric part, under assumption (A_{corr}). See Gannaz (2007b) for a more argued discussion on the influence of the linear part on estimation, explained through a parallel established with robust M-estimates.

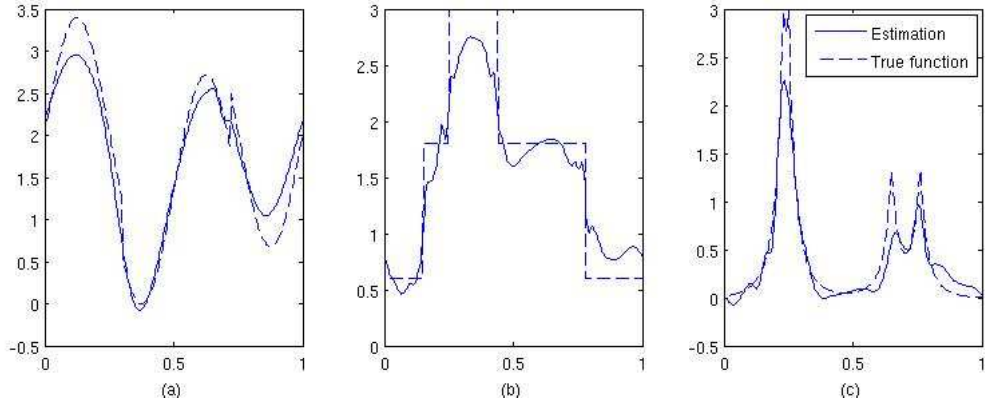


Figure 6: An example of estimation of the nonparametric part in Example 1, with the function *Sinus* in Figure (a), the function *Blocs* in Figure (b) and the function *Pics* in Figure (c).

2.5.3 Example 2: Binomial distribution

In Example 2, we consider a Binomial distribution: observations Y_i are such that $Y_i \times m$ are independently drawn from Binomial distributions $\mathcal{B}(\mu_i, m)$ with the parameter m equal to $m = \phi^{-1}$. The link function considered is the logistic. The mean is thus equal to

$$\mu_i = \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta}_0 + f_0(t_i))}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta}_0 + f_0(t_i))}.$$

The logistic link makes sense if the canonical parameter $\eta(\cdot)$ belongs to the interval $[-4, 4]$, as one can see for example page 28 of Fahrmeir & Tutz (1994). Consequently, to get a SNR of order 9 one may choose a parameter m in the binomial distribution equal approximatively to 200. Note that Binomial distribution is often chosen to classify the data. A number of 200 classes is not adapted for real data applications ; actually, a classical number of classes is 2, 3 or 4, which will lead to small SNRs and thus to a bad quality of estimation.

Due to this remark, we choose here to make a compromise and to apply the algorithm with a parameter m equal to 24, which corresponds to 25 classes. This choice is not coherent with a classification problem but allows to consider much reasonable SNRs. Indeed, with small values of m one is not in capacity of identifying the functional part in a GPLM. The observations of a simulated sample are represented in Figure 7. The results for this example are summarized in Table 4.

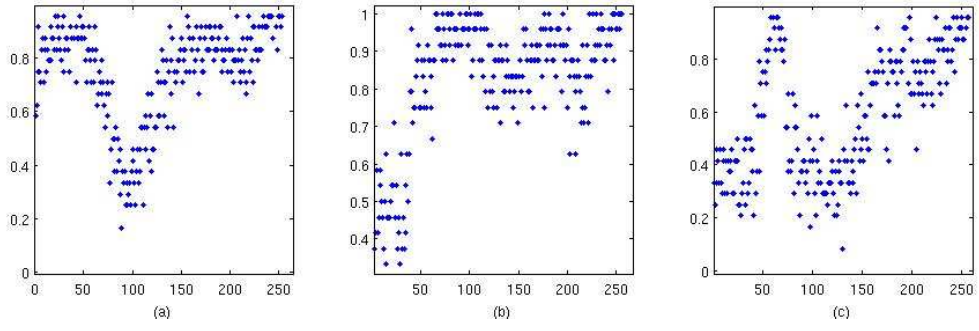


Figure 7: An example of the observations obtained on a simulation in Example 2, with the function *Sinus* in Figure(a), the function *Blocs* in Figure (b) and the function *Pics* in Figure (c).

Function	Estimation of β		Estimation of f		Time
	Mean SNR	Mean estimation	Mean SNR	MISE	
<i>Sinus</i>	2.1	0.8559(0.0597)	3.0	0.7656	17.0(3922)
<i>Blocs</i>	2.0	0.8581(0.0607)	2.9	0.7285	18.1(4142)
<i>Pics</i>	1.9	0.9487(0.0426)	1.4	0.4831	9.8(1886)

Table 4: Measures of quality the estimates over the 500 simulations in Example 2 with $n = 2^8$. For the parameter β , the true value is 1 and the value given is the mean value and standard deviation appears in brackets. In the column Time, the mean of numbers of iterations in the algorithm is given in brackets.

The bad quality of estimation is due to the bad SNRs of the model. In order to better analyse the results, we simulate a GPLM with a Gaussian distribution and the same SNRs. This was done by modifications on the dispersion parameter or choosing proportional covariates in the linear part. We obtain results given in Table 5.

Comparing Table 4 and Table 5, it appears that the estimation quality is better with a Gaussian framework than with a Binomial. Yet, the estimates in the Binomial GPLM seems satisfying for such SNRs. Figure 8 confirms that the estimate of the functional part only renders the shape of the function but does not allow to recover the function. Again this is in coherence with the small SNR.

Function	Estimation of β		Estimation of f		Time
	Mean SNR	Mean estimation	Mean SNR	MISE	
<i>Sinus</i>	2.3	1.0008(0.0333)	3.0	0.4219	1.04(253)
<i>Blocs</i>	1.9	1.0033(0.0373)	2.9	0.4547	1.87(451)
<i>Pics</i>	2.2	1.0002(0.0296)	1.6	0.3367	0.64(153)

Table 5: Measures of quality the estimates over the 500 simulations with a Gaussian distribution with similar SNRs than in Example 2 with $n = 2^8$. For the parameter β , the true value is 1 and the value given is the mean value and standard deviation appears in brackets. In the column Time, the mean of numbers of iterations in the algorithm is given in brackets.

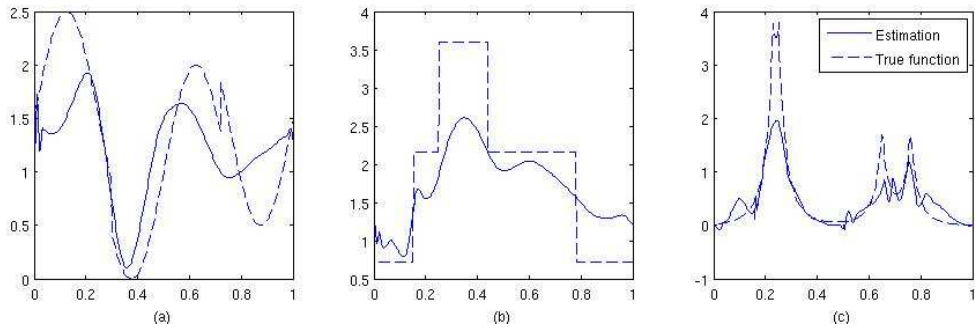


Figure 8: An example of estimation of the nonparametric part in Example 2, with the function *Sinus* in Figure(a), the function *Blocs* in Figure (b) and the function *Pics* in Figure (c).

When comparing the time costs, one can see that the algorithm is faster in a Gaussian framework. The reason is the fact that each step of the *backfitting* algorithm described above are exact calculation in a Gaussian case. While with a Binomial distribution each of these steps has to be solved iteratively. The number of iterations necessary to stabilize the algorithm is thus higher with the Binomial distribution.

This can also explain the lower quality of estimation in the Binomial distribution : the mean number of iterations in the algorithm is higher than 3900 with the *Sinus* and the *Blocs* functions, where the difference with the Gaussian distribution is the most important. Recall the maximal number of iterations is 5000. Thus perhaps sometimes the algorithm is not stabilized when it stops, explaining the lower quality.

2.5.4 Example 3: Poisson distribution

In Example 3, we consider a Poisson distribution: $y_i \sim \mathcal{P}(\mu_i)$ with $\mu_i = \exp(\eta_i)$. Observations of a simulated data sample are represented in Figure 9. The results for this example are summarized in Table 6.

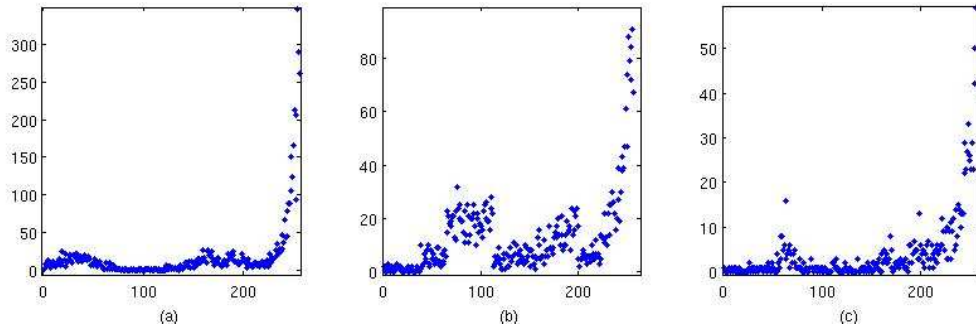


Figure 9: An example of the observations obtained on a simulation in Example 3, with the function *Sinus* in Figure(a), the function *Blocs* in Figure (b) and the function *Pics* in Figure (c).

Function	Estimation of β		Estimation of f		Time
	Mean SNR	Mean estimation	Mean SNR	MISE	
<i>Sinus</i>	9.1	1.0971(0.0763)	7.9	0.5548	11.4(2701)
<i>Blocs</i>	3.1	1.2020(0.1201)	8.2	0.4156	18.8(4479)
<i>Pics</i>	3.0	1.1876(0.1751)	7.1	0.4809	19.6(4528)

Table 6: Measures of quality the estimates over the 500 simulations in Example 3 with $n = 2^8$. For the parameter β , the true value is 1 and the value given is the mean value and standard deviation appears in brackets. In the column Time, the mean of numbers of iterations in the algorithm is given in brackets.

Similarly to what was done in Example 2, we also simulate data sets with similar SNRs with a Gaussian distribution, in order to be able to compare the qualities of the estimates to a given reference. Results obtained in a Gaussian case are given in Table 7.

Function	Estimation of β		Estimation of f		Time
	Mean SNR	Mean estimation	Mean SNR	MISE	
<i>Sinus</i>	9.3	0.9990(0.0101)	7.9	0.1875	0.61(147)
<i>Blocs</i>	3.2	0.9943(0.0311)	8.2	0.2261	1.39(336)
<i>Pics</i>	3.0	0.9957(0.0332)	7.1	0.1416	0.74(178)

Table 7: Measures of quality the estimates over the 500 simulations with a Gaussian distribution with similar SNRs than in Example 3 with $n = 2^8$. For the parameter β , the true value is 1 and the value given is the mean value and standard deviation appears in brackets. In the column Time, the mean of numbers of iterations in the algorithm is given in brackets.

The quality of estimation seems good, even if lower than in the Gaussian distribution framework. Like for the Binomial distribution, each step of the *backfitting* algorithm is solved by Fisher-scoring. Consequently, the time of calculation is higher than in the Gaussian model. Like with the Binomial distribution, the lowest qualities are observed when the mean of iteration numbers is high. Clearly, the fact that the mean of iteration numbers is more than 4400, with still a maximum at 5000, means that the algorithm often would have need more iterations to get stabilized. Thus the quality should probably be better with more iterations.

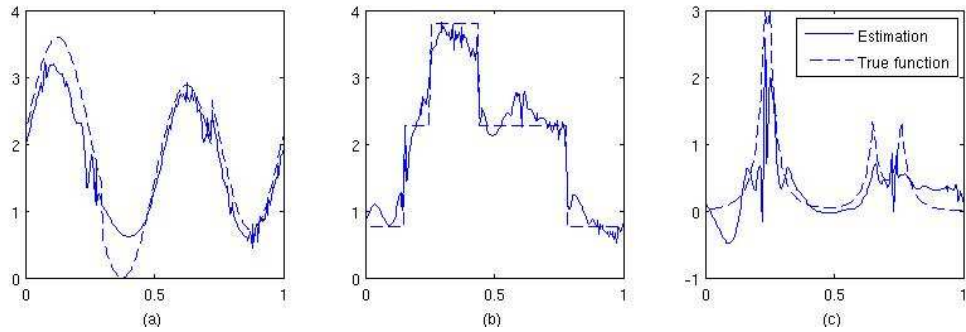


Figure 10: An example of estimation of the nonparametric part in Example 3, with the function *Sinus* in Figure (a), the function *Blocs* in Figure (b) and the function *Pics* in Figure (c).

Figure 10 gives an example of the estimation of the functional part obtained. One can see that visually the quality is satisfying. Note that one cannot compare with the Binomial case because of the difference in SNRs.

Conclusion

This paper proposes a penalized loglikelihood estimation in generalized partially linear models. With appropriate penalties, we establish the asymptotic near-minimaxity of the estimators for both the linear and the nonparametric parts of the model. The result holds for a large class of functions, possibly non smooth and the conditions of correlation between the covariate design of the linear part and the functional part are similar to literature's. Moreover with an ℓ^1 penalty on the wavelet coefficients of the function (leading to soft thresholding), the procedure appears to be adaptive relatively to the smoothness of the function. In this particular case we develop an implementation and observed satisfactory results on simulation studies.

Our ongoing research may deal with the estimation of the dispersion parameter in generalized partially linear models. Developments in non-equidistant designs for the nonparametric part should also be explored.

Acknowledgements

Pr. Antoniadis is gratefully acknowledged for his constructive comments and fruitful discussions. The author would also like to thank Pr. Gadat for giving some helpful references.

References

- Antoniadis, A. (1996). Smoothing noisy data with tapered coiflets series. *Scandinavian Journal of Statistics*, **23**(3), 313–330.
- Antoniadis, A., Besbeas, P., & Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with cubic variance functions. *Sankhyā: The Indian Journal of Statistics*, **363**(3), 309–327.
- Antoniadis, A., & Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, **96**(455), 939–967.

- Antoniadis, A., Gijbels, I., & Nikolova, M. (2009). Penalized likelihood regression for generalized linear models with nonquadratic penalties. *Ann. Inst. Stat. Math.*
- Antoniadis, A., & Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika*, **88**, 805–820.
- Bai, Z., Rao, C., & Wu, Y. (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica*, 237–254.
- Birgé, L., & Massart, P. (2000). An adaptive compression algorithm in Besov spaces. *Journ. Constr. Approx.*(16), 1–36.
- Boente, G., He, X., & Zhou, J. (2006). Robust estimates in generalized partially linear models. *The Annals of Statistics*, **34**(6), 2856–2878.
- Brown, L., Cai, T., & Zhou, H. (2008). Nonparametric regression in exponential families. Tech. Rep..
- Chang, X.-W., & Qu, L. (2004). Wavelet estimation of partially linear models. *Computational Statistics and Data Analysis*.
- Chen, H. (1987). Estimation of semiparametric generalized linear models. Tech. Rep.. State University of New York.
- Daubechies, I. (1992). *Ten lectures on wavelets*. (Vol. 61). SIAM press.
- Donoho, D., & Johnstone, I. (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics*, **26**(3), 879–921.
- Donoho, D., Johnstone, I., Kerkycharian, G., & Picard, D. (1995). Wavelet shrinkage: asymptotia? *Journal of Royal Statistics Society*, **57**(2), 301–369.
- Fadili, J., & Bullmore, E. (2005). Penalized partially linear models using sparse representation with an application to fMRI time series. *IEEE Transactions on Signal Processing*, **53**(9), 3436–3448.
- Fahrmeir, L., & Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*. Springer.
- Gannaz, I. (2007a). Estimation par ondelettes dans les modèles partiellement linéaires. Unpublished doctoral dissertation, Université Grenoble 1.

- Gannaz, I. (2007b). Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing*, **4**(17), 293–310.
- Huber, P. (1981). *Robust statistics*. Wiley Series in Probability and Mathematical Statistics.
- Loubes, J., & Van der Geer, S. (2002). Adaptive estimation with soft thresholding type penalties. *Statistica Neerlandica*, **56**(4), 454–479.
- Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(7), 674–693.
- Mallat, S. (1999). *A wavelet tour on signal processing*. (2 ed.). Academic Press.
- Mammen, E., & Van der Geer, S. (1997). Penalized quadi-likelihood estimation in partial linear models. *The Annals of Statistics*, **25**(3), 1014–1035.
- McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. (2 ed.). Chapman&Hall.
- Meyer, F. (2003). Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series. *IEEE transactions on medical imaging*, **22**, 315–324.
- Meyer, Y. (1992). *Wavelets and operators*. Cambridge University Press.
- Müller, M. (2001). Estimating and testing in generalized partial linear models – A comparative study. *Statistics and Computing*(11), 299–309.
- Reynaud-Bouret, P., & Rivoirard, V. (2010). Near optimal thresholding estimation of a Poisson intensity on the real line. *Electronic Journal of Statistics*, **4**, 172–238.
- Rice, J. (1986). Convergence rates for partially splined models. *Statistics and Probability Letters*, **4**, 203–208.
- Rockafellar, R. (1970). *Convex analysis*. Princeton University Press.
- Sardy, S., Antoniadis, A., & Tseng, P. (2004). Automatic smoothing with wavelets. *Journal of computational and graphical statistics*, **13**(2), 1–23.
- Severini, T., & Staniswalis, J. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of American Statistical Association*(89), 501–511.

- Severini, T., & Wong, W. (1992). Generalized profile likelihood and conditionally parametric models. *Annals of statistics*(20), 1768–1802.
- Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of Royal Statistical Society*, 50(3), 413–436.
- Van der Geer, S. (2000). *Empirical processes in M-estimation*. Cambridge University Press.

3 Appendix

The proof is structured as follows: in Section A, we justify the fact that under a structure (A_{corr}) of the covariates, with a well-chosen penalty, the scale part of the linear part does not intervene in the quality of the estimators. Section B studies the behaviour of the nonparametric part f , while Section C presents the asymptotic properties of the estimator for the linear regressor β .

A Decorrelation of the scale components

We suppose here that the assumption (A_{corr}) holds and that in the penalty only intervene the wavelets coefficients. To be more precise, given a vector of real values $\mathbf{e} = (e_1, \dots, e_n)^T$, the discrete wavelet transform of \mathbf{e} is given by a $n \times 1$ vector comprising both discrete scaling coefficients, noted $(\theta_{js,k}^S)_{k \in \mathbb{Z}}$, and discrete wavelet coefficients, $(\theta_{j,k}^W)_{j \geq j_S, k \in \mathbb{Z}}$. The wavelet inverse transform of the scaling coefficients $(\theta_{js,k}^S)_{k \in \mathbb{Z}}$ (fixing the wavelet coefficients equal to zeros) is noted e^S and the wavelet inverse transform of the wavelets coefficients $(\theta_{js,k}^W)_{k \in \mathbb{Z}}$ (fixing the scale coefficients equal to zeros) is noted e^W . The penalty is in this section assumed to be such that $Pen(f)$ is a function of f^W only. To better apprehend the mechanisms, let us decompose each components \mathbf{X}_i and f in the criterion into their scale parts \mathbf{X}_i^S and f^S and their wavelets parts \mathbf{X}_i^W and f^W .

Recall the estimation procedure consists in maximizing the criterion given in equation (5), first with respect to the functional part f and afterwards with respect to the linear component. For any

function f and any p -dimensional vector β , the criterion $K_n(f, \beta)$ given in (5) can be written

$$K_n(f, \beta) = \sum_{i=1}^n \ell \left(y_i, \mathbf{X}_i^{WT} \beta + a(f, X^S \beta)(t_i) \right) - \text{Pen}(a(f, X)) =: L_n(a(f, X^S \beta), X^W \beta), \quad (6)$$

with $a(f, X^S \beta)(t_i) = \mathbf{X}_i^{ST} \beta + f(t_i)$. It is worthy noticing the criterion $L_n(h, X^W \beta)$ does not depend of X^S for a given function h . For a fixed vector β , $\tilde{f}_\beta = \underset{f}{\operatorname{argmax}} K_n(f, \beta) = h_\beta - X^{ST} \beta$ with $h_\beta = \underset{h}{\operatorname{argmax}} L_n(h, \beta)$. Replacing f by its estimation, $K_n(\tilde{f}_\beta, \beta) = \sum_{i=1}^n \ell \left(y_i, \mathbf{X}_i^{WT} \beta + h_\beta(t_i) \right) - \text{Pen}(h_\beta)$. Consequently, the estimate $\hat{\beta}_n = \underset{\beta}{\operatorname{argmax}} K_n(\tilde{f}_\beta, \beta)$ is in fact defined independently from X^S in such a framework.

It is clear that the properties established for the argument maximizing the criterion K_n are available when maximizing the criterion L_n . When the structure of the covariates and of the penalty are well-adapted we will then consider without loss of generality that the covariates $(\mathbf{X}_i)_{i=1, \dots, n}$ contain only the wavelets components $(\mathbf{X}_i^W)_{i=1, \dots, n}$.

B Estimation of the nonparametric part by \tilde{f}_β

In this Section, we have to distinguish between the non-adaptive type penalties of Section 2.3.1 and the ℓ^1 penalty on wavelet coefficients. The whole scheme is the same, so the second case will be less detailed.

B.1 Nonadaptive case

Here, the penalty is supposed directly linked with an entropy control through assumptions (A5) and (A6).

B.1.1 A Taylor expansion

The difference between the criterions $K_n(f, \beta)$ and $K_n(f_0, \beta)$ is equal to:

$$\begin{aligned} \frac{1}{n} (K_n(f, \beta) - K_n(f_0, \beta)) = \\ \frac{1}{n} \sum_{i=1}^n \ell(y_i, \mathbf{X}_i^T \beta + f(t_i)) - \sum_{i=1}^n \ell(y_i, \mathbf{X}_i^T \beta_0 + f_0(t_i)) - \frac{\lambda}{n} (Pen(f) - Pen(f_0)). \end{aligned}$$

Note $\eta_{0,i} = \mathbf{X}_i^T \beta_0 + f_0(t_i)$. A Taylor expansion of degree 2 at the points $(\ell(y_i, \eta_{0,i}))_i$ gives:

$$\begin{aligned} \frac{1}{n} (K_n(f, \beta) - K_n(f_0, \beta)) \\ = \underbrace{\frac{1}{n} \sum_{i=1}^n \dot{\ell}(y_i, \eta_{0,i}) (f(t_i) - f_0(t_i))}_{A_1} - \underbrace{1/2 \frac{1}{n} \sum_{i=1}^n \ddot{\ell}(\eta_{0,i}) (f(t_i) - f_0(t_i))^2}_{A_2} \\ - \underbrace{\frac{1}{n} \sum_{i=1}^n \ddot{\ell}(\eta_{0,i}) (\mathbf{X}_i^T (\beta - \beta_0)) (f(t_i) - f_0(t_i))}_{Co} - \frac{\lambda}{n} (Pen(f) - Pen(f_0)) + T_1(f, \beta) \end{aligned}$$

$$\text{with } T_1(f, \beta) \leq \frac{1}{6} \sup_i \sup_{\eta} \ddot{\ell}(\eta_i) \frac{1}{n} \sum_i \left(\left| \mathbf{X}_i^T (\beta - \beta_0) + f(t_i) - f_0(t_i) \right|^3 + |f(t_i) - f_0(t_i)|^3 \right).$$

Let us study the term $T_1(f, \beta)$. Under assumption (A3),

$$T_1(f, \beta) \leq cste \frac{1}{n} \sum \left(\left| \mathbf{X}_i^T (\beta - \beta_0) + f(t_i) - f_0(t_i) \right|^3 + \left| \mathbf{X}_i^T (\beta - \beta_0) \right|^3 \right).$$

Using the convexity of $x \mapsto x^3$, it comes:

$$T_1(f, \beta) \leq cste \left(\underbrace{\frac{1}{n} \sum_{i=1}^n \left| \mathbf{X}_i^T (\beta - \beta_0) \right|^3}_{T_{11}} + \underbrace{\frac{1}{n} \sum_{i=1}^n |f(t_i) - f_0(t_i)|^3}_{T_{12}} \right).$$

Then, one can easily see that when assumptions (A1) and (A2) hold, $T_{11} = o(p/n)$ and that $T_{12} = \frac{1}{n} \sum |f(t_i) - f_0(t_i)|^3$. Consequently, if pv_n^2/n is bounded then for every function f satisfying $\frac{v_n^2}{n} \sum_i |f(t_i) - f_0(t_i)|^3 \leq c$ we have

$$v_n^2 T_1 \leq C.$$

When $v_n = n^{1/(2+\nu)}$ with $\nu > 0$, the sequence pv_n^2/n is bounded if $pn^{-\nu/(2+\nu)}$ is bounded.

Let

$$S_n(f) = \frac{v_n^2}{n} (K_n(f, \beta) - K_n(f_0, \beta) - A_1 + A_2 - Co) + \frac{\lambda v_n^2}{n} (Pen(f) - Pen(f_0)).$$

For every sequence $u_n \rightarrow 0$ and for all function f satisfying $\frac{v_n^2}{n} \sum_i |f(t_i) - f_0(t_i)|^3 \leq c$, we prove $u_n S_n(f) \rightarrow 0$. Fix u_n a decreasing sequence going to 0. Convexity argumentation (Theorem 10.8 in Rockafellar (1970) with a change of variable on f to obtain a set independent from n) gives:

$\sup_{\frac{v_n^2}{n} \sum_i |f(t_i) - f_0(t_i)|^3 \leq c} u_n S_n(f) \rightarrow 0$. Consequently:

$$\sup_{N(f) \leq c} u_n S_n(f) \rightarrow 0, \text{ where } N(f) = v_n^2 \|f - f_0\|_n^2 + J(f), \quad (7)$$

noticing that $\frac{v_n^2}{n} \sum_i |f(t_i) - f_0(t_i)|^3 \leq v_n^2 \|f - f_0\|_n^2 \|f - f_0\|_\infty$, and that we consider the maximisation problem on the set $\{f, \|f\|_\infty \leq C_\infty\}$.

Remark: We need $f \mapsto S_n(f/v_n)$ to be concave for the norm $\|\cdot - f_0\|_n$ for every n . It is worthy noticing the penalty Pen does not intervene in the definition of S_n .

B.1.2 Behaviour of the different terms

Similarly to Bai et al. (1992), we suppose $N(f) \geq c_n$ with $c_n u_n \rightarrow \infty$. Then one can build a sequence c'_n satisfying $c'_n u_n \rightarrow \infty$ et $c'_n \leq c_n$ and such that $\sup_{N(f) \leq c'_k} u_k S_k$ goes to 0 when k goes to infinity.

For $N(f) = c'_n$, this means:

$$\frac{v_n^2}{n} (K_n(f, \beta) - K_n(f_0, \beta) - A_1 + A_2 - Co) + \frac{\lambda v_n^2}{n} (Pen(f) - Pen(f_0)) = \mathcal{O}(1). \quad (8)$$

- **Control of the term** $A_1 = \frac{1}{n} \sum_{i=1}^n \dot{\ell}(y_i, \eta_{0,i}) (f(t_i) - f_0(t_i))$.

Following Mammen & Van der Geer (1997), let \mathcal{A} be a set of uniformly bounded functions on $[0, 1]$ satisfying

$$\limsup_{n \rightarrow \infty} \sup_{\delta > 0} \delta^\nu \mathcal{H}(\delta, \mathcal{A}, \|\cdot\|_n) < \infty.$$

Suppose $\frac{f}{1+J(f)} \in \mathcal{A}$. Then, similarly to Mammen & Van der Geer (1997), when assumption (A4.1) holds:

$$\sup_{f, \|f\|_\infty \leq C_\infty} \sqrt{n} \frac{A_1}{\left(\frac{\|f - f_0\|_n}{1+J(f)} \vee n^{-1/(2+\nu)} \right)^{1-\nu/2}} = \mathcal{O}_{\mathbb{P}}(1).$$

Under assumption (A6), if $N(f) = c'_n$, then $J(f) \leq c'_n$, and consequently, we obtain that with $v_n = n^{1/(2+\nu)}$:

$$\sup_{f, \|f\|_\infty \leq C_\infty, N(f) \leq c'_n} v_n^2 A_1 = o_{\mathbb{P}}(c'_n).$$

- **Control of the term** $A_2 = 1/2 \frac{1}{n} \sum_{i=1}^n \ddot{b}(\eta_{0,i}) (f(t_i) - f_0(t_i))^2$.

Thanks to assumption (A3) we immediately get: $v_n^2 A_2 \geq \ddot{b}_0 (v_n \|f - f_0\|_n)^2$ and thus $\inf_{N(f)=c'_n} A_2 \geq \ddot{b}_0 c_n'^2$.

- **Control of the term** $C_0 = \frac{1}{n} \sum_{i=1}^n \ddot{b}(\eta_{0,i}) ((\beta - \beta_0)^T \mathbf{X}_i) (f(t_i) - f_0(t_i))$.

Cauchy-Schwarz inequality applied to C_0 gives:

$$C_0 \leq \sup_i \ddot{b}(\eta_{0,i}) \|f(t_i) - f_0(t_i)\|_n \left(\frac{1}{n} \sum \|\mathbf{X}_i\|^2 \right)^{1/2} \|\beta - \beta_0\|.$$

When $\sqrt{n} \|\beta - \beta_0\| \leq c$ and assumptions (A1) to (A3) hold, we get:

$$C_0 \leq C(\beta) n^{-1/2} \|f - f_0\|_n,$$

with $C(\beta)$ independent from f . Thus:

$$\sup_{N(f) \leq c'_n} v_n^2 C_0 = o_{p.s.}(c'_n),$$

for $v_n = n^{\frac{1}{1+2\nu}}$.

B.1.3 Conclusion

Using together the convergence (7), the bounds of A_1 , A_2 and C_0 and assumption (A5), we obtain:

$$\sup_{\{f, N(f)=c'_n\}} u_n \frac{v_n^2}{n} (K_n(f, \beta) - K_n(f_0, \beta)) < 0$$

with a probability going to 1 when n goes to infinity. Concavity of assumption (A6) (because it implies the decrease of the slopes) allows to extend this result to

$$\sup_{\{f, N(f) \geq c'_n\}} u_n \frac{v_n^2}{n} (K_n(f, \beta) - K_n(f_0, \beta)) < 0.$$

The estimator \tilde{f}_β is the argument realizing the maximum of $K_n(\cdot, \beta)$ and so $\mathbb{P}(N(f) \geq c'_n) \rightarrow 0$.

B.2 Adaptive case, with an ℓ^1 penalty

In this section $Pen(f) = \sum_{i=i_s}^n |\theta_i^W|$ with θ^W vector of the wavelet coefficients of f . In the first time, we are willing to study how this penalty can lead to a functional space for which we can control the entropy as above. The underlying idea is to distinguish the behaviour of the penalty among the resolution degree of the wavelets coefficients. We introduce $i_W = \left(\frac{n}{\log(n)}\right)^{1/(1+2s)}$. If the resolution level is higher we will establish a link with an other penalty, but if the resolution level is smaller, then we are going to see that the norm $\|\cdot\|_n$ offers a sufficient control.

B.2.1 Highest resolution levels

Let $Pen_{i_W}(f)$ be $Pen_{i_W}(f) = \sum_{i=i_W}^n |\theta_i^W|$.

Minoration

We aim to bound above the truncated penalty $Pen_{i_W}(f)$ using $J_{i_W}(f) = n^{-\rho/2} \sum_{i=i_W}^n |\theta_{j,k}^W|^\rho$ with $\rho = 2/(1+2s)$. To this objective, we decompose as follows :

$$Pen_{i_W}(f) = \sum_{i \geq i_W, |\theta_i^W| < \varepsilon} |\theta_i^W| + \sum_{i \geq i_W, |\theta_i^W| \geq \varepsilon} |\theta_i^W|.$$

- Hölder inequalities gives:

$$\sum_{|\theta_i^W| \leq \varepsilon} |\theta_i^W|^\rho \leq \#\{|\theta_i^W| \leq \varepsilon\}^{1-\rho} \left(\sum_{|\theta_i^W| \leq \varepsilon} |\theta_i^W| \right)^\rho.$$

Observing that $\#\{|\theta_{j,k}^W| \leq \varepsilon\} = \sum \mathbb{1}_{|\theta_i|/\lambda \leq 1} \geq \varepsilon^{-\rho} \sum_{|\theta_i^W| \leq \varepsilon} |\theta_i^W|^\rho$, we have

$$\left(\sum_{|\theta_i^W| \leq \varepsilon} |\theta_i^W| \right) \geq \varepsilon^{1-\rho} \left(\sum_{|\theta_i^W| \leq \varepsilon} |\theta_i^W|^\rho \right).$$

- The bound $\sum_{|\theta_i^W| \geq \varepsilon} |\theta_i^W| \geq \varepsilon^{1-\rho} \sum_{|\theta_i^W| \geq \varepsilon} |\theta_i^W|^\rho$ is evident.

It follows that: $Pen_{i_W}(f) \geq \varepsilon^{1-\rho} n^{\rho/2} J_{i_W}(f)$. Taking $\varepsilon = \lambda$, $v_n = \left(\frac{\log(n)}{n}\right)^{2s/(1+2s)}$ and $\lambda \geq \sqrt{\log(n)}$, we obtain that $\frac{\lambda v_n^2}{n} Pen_{i_W}(f) \geq J_{i_W}(f)$.

Majoration

We are now willing to bound $Pen_{i_W}(f_0)$. Using Hölder inequality,

$$\sum_k |\theta_{0,jk}^W| \leq 2^{j(1-1/\pi)} \left(\sum_k |\theta_{0,jk}^W|^\pi \right)^{1/\pi}$$

because as ψ and f admits compact supports the number of non-zero coefficients at a resolution level j is equivalent to 2^j . Thus, $Pen_{i_W}(f_0) \leq \sum_{j \geq j_W} 2^{-j(s-1/2)} n^{1/2} \|f_0\|_{s,\pi,\infty}$, and

$$\frac{\lambda v_n^2}{n} Pen_{i_W}(f_0) \leq \frac{\lambda v_n^2}{n^{1/2}} i_W^{-(s-1/2)} \|f_0\|_{s,\pi,\infty}.$$

For $v_n = \left(\frac{\log(n)}{n}\right)^{s/(1+2s)}$ and $\lambda \sim \sqrt{\log(n)}$, it is sufficient that $i_W \geq \left(\frac{n}{\log(n)}\right)^{1/(1+2s)}$ to get $\frac{\lambda v_n^2}{n} Pen_{i_W}(f_0) \leq \|f_0\|_{s,\pi,\infty}$.

B.2.2 Lowest resolution levels

We are interested in the study of $Pen_{i_S}(f) = \sum_{i=i_S}^{i_W} |\theta_i^W|$. Note that $|Pen_{i_S}(f) - Pen_{i_S}(f_0)| \leq \sum_{i=i_S}^{i_W} ||\theta_i^W| - |\theta_{0,i}^W|| \leq \sum_{i=i_S}^{i_W} |\theta_i^W - \theta_{0,i}^W|$. Using Cauchy-Schwarz inequality,

$$|Pen_{i_S}(f) - Pen_{i_S}(f_0)| \leq i_W^{1/2} \left(\sum_{i=i_S}^{i_W} |\theta_i^W - \theta_{0,i}^W|^2 \right)^{1/2} = (ni_W)^{1/2} \|f - f_0\|_n.$$

For $v_n = \left(\frac{\log(n)}{n}\right)^{s/(1+2s)}$ and $\lambda = \sqrt{\log(n)}$, it is sufficient that $i_W \leq \left(\frac{n}{\log(n)}\right)^{1/(1+2s)}$ to deduce $\frac{\lambda v_n^2}{n} |Pen_{i_S}(f) - Pen_{i_S}(f_0)| \leq v_n \|f - f_0\|_n$.

To conclude the penalty study, we have established that, with the adapted choice of i_W ,

$$\frac{\lambda v_n^2}{n} (Pen(f) - Pen(f_0)) - R \geq J_{i_W}(f) \quad \text{with} \quad |R| \leq \|f_0\|_{s,\pi,\infty} + v_n \|f - f_0\|_n.$$

B.2.3 Bias term $A_1 = \frac{1}{n} \sum_{i=1}^n \dot{\ell}(y_i, \eta_{0,i}) (f(t_i) - f_0(t_i))$.

Let $N(\cdot)$ and $J(\cdot)$ be defined as follows : $N(f) = v_n^2 \|f - f_0\|_n^2 + J_{i_W}(f)$ and $J(f) = n^{-\rho/2} \sum_{i=i_S}^n |\theta_i^W|^\rho$. We are going to put in evidence the link between these two quantities. By Cauchy-Schwarz inequality,

$$\sum_{i=i_S}^{i_W} |\theta_i^W - \theta_{0,i}^W|^\rho \leq i_W^{1-\rho/2} \left(\sum |\theta_i^W - \theta_{0,i}^W|^2 \right)^{\rho/2} = i_W^{1-\rho/2} n^{\rho/2} \|f - f_0\|_n^\rho.$$

Then for $v_n = \left(\frac{\log(n)}{n} \right)^{2s/(1+2s)}$, as $i_W \leq (n/\log(n))^{1/(1+2s)}$, we have: $\sum_{i=i_S}^{i_W} |\theta_i^W - \theta_{0,i}^W|^\rho \leq v_n^\rho \|f - f_0\|_n^\rho$. When $v_n^2 \|f - f_0\|^2 = c'_n \rightarrow \infty$, it implies that $J(f - f_0) \leq N(f) + J_{i_W}(f_0)$.

We will consider the functional set $\mathcal{A} = \{f, J(f) \leq C_0\}$. Lemme 4.3. of Loubes & Van der Geer (2002) states the entropy of \mathcal{A} satisfy $\mathcal{H}(\mathcal{A}, \delta, \|\cdot\|_n) \leq A \delta^{-\frac{2\rho}{2-\rho}} (\log(n) + \log(1/\delta))$ where A is a constant. With $\nu = \frac{2\rho}{2-\rho} = 1/s$, this can be written $\int_{1/n}^R \mathcal{H}(\mathcal{A}, u, \|\cdot\|_n)^{1/2} du \leq A \sqrt{\log(n)} R^{1-\nu/2}$.

Corollary 8.3 of Van der Geer (2000) implies that under assumption (A4.2), for $R > \delta/\sigma > 1/n$, and $C > 1$, we have:

$$\mathbb{P} \left(\sup_{f \in \{g, g \in \mathcal{A}, \|g\|_n \leq R\}} \frac{1}{\sqrt{\log(n)} \sqrt{n}} \sum_{i=1}^n \dot{\ell}(y_i, \eta_{0,i}) f(t_i) \geq 2AC_0 \delta^{1-\nu/2} C \right) \leq C_0 \exp(-A_0^2 \delta^{-\nu} C).$$

Following the proof of Lemma 8.4 in Van der Geer (2000) we can deduce the existence of a constant c depending only of A, ν, R and of constants in (A4.2) such that:

$$\text{forall } T \geq c, \mathbb{P} \left(\sup_{f \in \{g \in \mathcal{A}, \|g\|_n \leq R\}} \frac{\sqrt{\log(n)} \sum_{i=1}^n \dot{\ell}(y_i, \eta_{0,i}) f(t_i)}{\sqrt{n} \|g\|_n^{1-\nu/2}} \geq T \right) \leq c \exp(-T^2/c^2).$$

Assume $g = \frac{f-f_0}{N(f)+J_{i_W}(f_0)}$. According to previous developments, g belongs to \mathcal{A} and $\|g\|_n \leq 1$. Thus,

$$\sup_f \sqrt{\log(n)} \sqrt{n} \frac{A_1}{N(f) + J_{i_W}(f_0)} = \mathcal{O}_{\mathbb{P}}(n^{-1/2} \sqrt{\log(n)} \left(\frac{\|f - f_0\|_n}{N(f) + J_{i_W}(f_0)} \right)^{1-\nu/2}).$$

With $v_n = \left(\frac{\log(n)}{n} \right)^{1/(2+\nu)}$,

$$\begin{aligned} \sup_{f, N(f)=c'_n} v_n^2 A_1 &= \mathcal{O}_{\mathbb{P}}(n^{-1/2} \sqrt{\log(n)} v_n^{1+\nu/2} c_n'^{1/2-\nu/4} c_n'^{\nu/2}) \\ &= \mathcal{O}_{\mathbb{P}}(c_n'^{1/2+\nu/4}) \end{aligned}$$

As $0 < \nu < 2$, immediately, $\sup_{f, N(f)=c'_n} v_n^2 A_1 = o_{\mathbb{P}}(c'_n)$ when $c'_n \rightarrow \infty$

B.2.4 End of the proof

The scheme is very similar to what has been presented in the nonadaptive case and will not be detailed here.

C Estimation of the linear part: study of the behaviour of $\widehat{\beta}_n$

With a Taylor expansion of order 2,

$$\begin{aligned}
 K(\tilde{f}_{\beta}, \beta) - K(\tilde{f}_{\beta}, \beta_0) &= \sum_{i=1}^n \ell(y_i, \mathbf{x}_i^T \beta + \tilde{f}_{\beta}(t_i)) - \sum_{i=1}^n \ell(y_i, \mathbf{x}_i^T \beta_0 + \tilde{f}_{\beta}(t_i)) \\
 &= \underbrace{\sum_{i=1}^n \dot{\ell}(y_i, \eta_{0,i}) \left(\mathbf{x}_i^T (\beta - \beta_0) \right)}_{B_1} - \underbrace{1/2 \sum_{i=1}^n \ddot{b}(\eta_{0,i}) \left(\mathbf{x}_i^T (\beta - \beta_0) \right)^2}_{B_2} \\
 &\quad - \underbrace{\sum_{i=1}^n \ddot{b}(\eta_{0,i}) \left(\mathbf{x}_i^T (\beta - \beta_0) \right) \left(\tilde{f}_{\beta}(t_i) - f_0(t_i) \right)}_{\tilde{C}_0} + T_2(\beta)
 \end{aligned}$$

C.1 Behaviour of the different terms

C.1.1 Control of \tilde{C}_0 .

We should first study the convergence of the rest term T_2 but the mechanisms which intervene in the majoration appears more clearly in \tilde{C}_0 . Write $\tilde{C}_0 = \|\tilde{f}_{\beta} - f_0\|_{\infty} \sum_{j=1}^p \tilde{C}_{0j}(\beta_j - \beta_{0,j})$ with

$$\tilde{C}_{0j} = \sum_{i=1}^n \ddot{b}(\eta_{0,i}) \bar{X}_{i,j} \frac{(\tilde{f}_{\beta}(t_i) - f_0(t_i))}{\|\tilde{f}_{\beta} - f_0\|_{\infty}}.$$

Without assumption (A_{corr})

Applying Cauchy-Schwarz leads to

$$\|\tilde{C}o_j\| \leq \|\tilde{f}_\beta - f_0\|_n \ddot{b}_\infty \left(\frac{1}{n} \sum X_{i,j}^2 \right)^{1/2}.$$

Under assumptions (A1), (A3) and what precedes, it comes $v_n \|\tilde{C}o\| = \mathcal{O}_{\mathbb{P}}(1)$.

With (A_{corr})

We assume the penalty does not deals with scale components of the nonparametric part of the model. As explained in Section A, we consider actually that for all $i = 1, \dots, n$, the covariate \mathbf{X}_i has a null scale representation. Assumption (A_{corr}) implies moreover that the polynomial functions g_j have a null wavelet representation and so $\tilde{C}o = \sum_{i=1}^n \ddot{b}(\eta_{0,i}) \left(\xi_i^{WT}(\beta - \beta_0) \right) \left(\tilde{f}_\beta(t_i) - f_0(t_i) \right)$, where ξ_i^W represents the wavelets part of ξ_i .

The wavelet transform is orthonormal and thus ξ_i^W has the same properties of ξ_i . When these variables satisfy an subgaussian assumption such as (A4.1) or (A4.2), we can apply a control of the term using entropy similar to what has been done before, lying on Van der Geer (2000).

When the entropy satisfies $\mathcal{H}(\mathcal{A}, \delta, \|\cdot\|_n) \leq A\delta^{-\nu}$, following Mammen & Van der Geer (1997) we get:

$$\sup_{\sqrt{n}\|\beta - \beta_0\| \leq c} \frac{\tilde{C}o_j}{\sqrt{n} \left(\frac{\|\tilde{f}_\beta - f_0\|_n}{\|\tilde{f}_\beta - f_0\|_\infty} \vee n^{-1/(2+\nu)} \right)^{1-\nu/2}} = \mathcal{O}_{\mathbb{P}}(1).$$

Using $\|\tilde{f}_\beta - f_0\|_n = \mathcal{O}_{\mathbb{P}}(v_n^{-1})$ for $\sqrt{n}\|\beta - \beta_0\| \leq c$, it comes:

$$\tilde{C}o = \mathcal{O}_{\mathbb{P}}(1) \left(v_n^{-1} \vee n^{-1/(2+\nu)} \right)^{1-\nu/2} \sqrt{n} \sum_{j=1}^p |\beta_j - \beta_{0,j}|.$$

Using Cauchy-Schwarz inequality:

$$\tilde{C}o = \mathcal{O}_{\mathbb{P}}(1) \left(v_n^{-1} \vee n^{-1/(2+\nu)} \right)^{1-\nu/2} \sqrt{np^{1/2}} \|\beta - \beta_0\|.$$

Provided $\rho_n = \left(v_n^{-1} \vee n^{-1/(2+\nu)} \right)^{1-\nu/2} p^{1/2} \rightarrow 0$, we have $\tilde{C}o = o_{\mathbb{P}}(1)$ for β such that $\sqrt{n}\|\beta - \beta_0\| \leq c$.

When the subgaussian assumption (A4.2) is weakened in the exponential tails assumption (A4.1) we can probably obtain a similar result using Lemma 8.4 of Van der Geer (2000).

With the ℓ^1 -penalty the majoration for the δ -entropy becomes $\mathcal{H}(\mathcal{A}, \delta, \|\cdot\|_n) \leq A\delta^{-1/s}(\log(n) + \log(1/\delta))$. Using again Van der Geer (2000) we have

$$\sup_{\sqrt{n}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq c} \tilde{C}o_j = \sqrt{n} \left(\frac{n}{\log(n)} \right)^{(2s-1)/(1+2s)},$$

and so

$$\tilde{C}o = \mathcal{O}_{\mathbb{P}}(1) \left(\frac{n}{\log(n)} \right)^{(2s-1)/(1+2s)} p^{1/2} \sqrt{n} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|.$$

It is sufficient to suppose $\rho_n^2 = \left(\frac{n}{\log(n)} \right)^{(s-1/2)/(1+2s)} p \rightarrow 0$ to get $\tilde{C}o = o_{\mathbb{P}}(1)$ for all $\sqrt{n}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq c$.

C.1.2 Control of T_2

The rest term in the Taylor expansion T_2 is bounded by:

$$T_2 \leq cste \sum \left| \left(\mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \tilde{f}_{\boldsymbol{\beta}}(t_i) - f_0(t_i) \right)^3 - \left(\tilde{f}_{\boldsymbol{\beta}}(t_i) - f_0(t_i) \right)^3 \right|.$$

We decompose in three terms:

- $T_{21} = \sum \left| \mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right|^2 |\tilde{f}_{\boldsymbol{\beta}}(t_i) - f_0(t_i)|$

$$T_{21} \leq \sup_i \|\mathbf{X}_i^T\| \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \sum |\mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)| |\tilde{f}_{\boldsymbol{\beta}}(t_i) - f_0(t_i)|.$$

Proceeding as for the term $\tilde{C}o$, when $\sqrt{n}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq c$, we obtain:

$$T_{21} \leq \sup_i \|\mathbf{X}_i^T\| \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \mathcal{O}_{\mathbb{P}}(\rho_n).$$

Assumption (A2) implies $T_{21} = o_{\mathbb{P}}(\rho_n)$.

- $T_{22} = \sum \left| \mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right| |\tilde{f}_{\boldsymbol{\beta}}(t_i) - f_0(t_i)|^2$. Note that we have

$$T_{22} \leq \|\tilde{f}_{\boldsymbol{\beta}} - f_0\|_{\infty} \sum |\mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)| |\tilde{f}_{\boldsymbol{\beta}}(t_i) - f_0(t_i)|.$$

And when $\sqrt{n}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq c$, the asymptotic behaviour is $T_{22} = \mathcal{O}_{\mathbb{P}}(\|\tilde{f}_{\boldsymbol{\beta}} - f_0\|_{\infty} \rho_n)$.

- $T_{23} = \sum \left| \mathbf{X}_i^T (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right|^3$. Under assumption (A1) and (A2), $T_{23} \leq o(1) (\sqrt{n} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_n)^3$.

We have proved that for all $\boldsymbol{\beta}$ satisfying $\sqrt{n} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_n \leq c_2$, the term T_2 goes to 0 when n goes to infinity, provided that $\rho_n \rightarrow 0$.

C.1.3 Control of $B_2 = 1/2 \sum_{i=1}^n \ddot{b}(\eta_{0,i}) \left(\mathbf{X}_i^T (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right)^2$.

Recall that assumption (A3) stands that $\sum_{i=1}^n \ddot{b}(\eta_{0,i}) \mathbf{X}_i^T \mathbf{X}_i$ goes to a strictly positive matrix Σ_0 when n goes to infinity. Thus, up to a constant c , we have $B_2 \geq c\gamma(\Sigma_0)n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2$, with $\gamma(\Sigma_0)$ smallest eigenvalue of Σ_0 . Consequently $\inf_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = c'_n} B_2 \geq c\gamma(\Sigma_0)nc_n'^2$.

C.1.4 Control of $B_1 = \sum_{i=1}^n \dot{\ell}(y_i, \eta_{0,i}) \left(\mathbf{X}_i^T (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right)$

Write $B_1 = B(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ with $B = \sum_{i=1}^n \dot{\ell}(y_i, \eta_{0,i}) \mathbf{X}_i^T$ an $n \times p$ -dimensional matrix. The norm of B satisfy the inequality:

$$\|B\|^2 \leq \left(\frac{1}{n} \sum \dot{\ell}(y_i, \eta_{0,i})^2 \right) \sum_{j=1}^p \left(\frac{1}{n} \sum ((\mathbf{X}_i^T)_j)^2 \right).$$

When assumption (A_{corr}) is satisfied, the law of large numbers implies that

$$\|B\|^2 \leq \phi \sum_{j=1}^p \sigma_j^2 o_{p.s.}(1).$$

Consequently

$$\sup_{\sqrt{n} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = c'_n} B_1 = \bigcirc_{\mathbb{P}}(c'_n).$$

C.2 Conclusion

We have proved that

$$R_n(\beta) := \left(K_n(\tilde{f}_\beta, \beta) - K_n(\tilde{f}_\beta, \beta_0) \right) - B_1 + B_2$$

goes to 0 for all β such that $\sqrt{n}\|\beta - \beta_0\| \leq c$. Using convexity (see Rockafellar (1970) as before) the convergence is available for the supremum $\sup_{\sqrt{n}\|\beta - \beta_0\| \leq c} R_n(\beta) \rightarrow 0$.

Suppose now that $\sqrt{n}\|\hat{\beta} - \beta_0\| \geq c_n$ with $c_n \rightarrow \infty$. Then we can build a sequence c'_n satisfying $c'_n \rightarrow \infty$, $c'_n \leq c_n$ and $\sup_{\sqrt{n}\|\beta - \beta_0\| = c'_n} R_n(\beta) \rightarrow 0$. Note that

$$R_n(\beta) = \left(K_n(\tilde{f}_\beta, \beta) - K_n(f_0, \beta_0) \right) - \left(K_n(\tilde{f}_\beta, \beta_0) - K_n(f_0, \beta_0) \right) - B_1 + B_2.$$

We have $K_n(\tilde{f}_\beta, \beta_0) - K_n(f_0, \beta_0) < 0$. Using moreover the studies of terms B_1 and B_2 show that

$$\sup_{\sqrt{n}\|\beta - \beta_0\| = c'_n} \left(K_n(\tilde{f}_\beta, \beta) - K_n(f_0, \beta_0) \right) \leq \mathcal{O}_{\mathbb{P}}(c'_n) - kc'_n{}^2,$$

with k strictly positive constant, and thus

$$\sup_{\sqrt{n}\|\beta - \beta_0\| = c'_n} \left(K_n(\tilde{f}_\beta, \beta) - K_n(f_0, \beta_0) \right) < 0$$

with a probability going to 1 when n goes to infinity.

Convexity gives:

$$\sup_{\sqrt{n}\|\beta - \beta_0\| \geq c'_n} K_n(\tilde{f}_\beta, \beta) - K_n(f_0, \beta_0) < 0$$

with a probability going to 1. As the estimator $\hat{\beta}_n$ minimizes $K_n(\tilde{f}_\beta, \beta)$, we conclude

$$\mathbb{P}(\sqrt{n}\|\hat{\beta}_n - \beta_0\| \geq c_n) \rightarrow 0.$$